CLASSIFICATION FACTORIELLE HIÉRARCHIQUE OPTIMISÉE DES LIGNES ET DES COLONNES D'UN TABLEAU DE CONTINGENCE

Jean-Jacques DENIMAL [1]

RÉSUMÉ

Etant donné un tableau de contingence k_{IJ}, deux classifications hiérarchiques sont construites indépendamment sur I et J selon un algorithme particulier où chaque nœud obtenu est issu d'une analyse des correspondances particulière. Un algorithme d'optimisation du type de celui des nuées dynamiques est ensuite appliqué aux classes de chacune des deux hiérarchies. Enfin, une procédure d'élagage des branches permet de se séparer des nœuds non significatifs. Les deux hiérarchies optimisées et élaguées sont ensuite interprétées mutuellement, chaque association significative étant révelée par un test conditionnel exact basé sur un modèle hypergéométrique. Un exemple d'application au tableau de contingence croisant départements et candidats à l'élection présidentielle de 1995 est ensuite mené.

Mots-clés : Tableau de contingence, Classification hiérarchique, Analyse des correspondances, Test conditionnel exact, Optimisation, Elagage.

ABSTRACT

Two hierarchical classifications are built on the sets I et J of a two-way contingency table k_{IJ}, using a new algorithm building each node from a particular correspondence analysis. In a second step, the classes of these two hierarchies are optimized through a type k-means procedure. Then, a pruning algorithm allows us to restrict the optimized trees to their significant nodes. Finally, the optimized and pruned hierarchies are mutually interpreted, each significant association being revealed through an exact conditional test based on the hypergeometric model. The methodogogy is then applied to the contingency table crossing departements and candidates to the 1995 presidential election.

Keywords : Contingency table, Hierarchical classification, Correspondence analysis, Exact conditional test, Optimization, Pruning techniques.

1. Université des Sciences et Technologies de Lille,
e-mail : jean-jacques.denimal@univ-lille1.fr

1. Introduction

L'objet de la méthode proposée est d'unifier dans une même approche, l'analyse des correspondances d'un tableau de contingence et les deux classifications hiérarchiques construites sur les lignes et les colonnes. Pour chacune de ces deux hiérarchies, une technique d'optimisation assure la qualité des classes obtenues et une technique d'élagage la significativité de celles-ci. Enfin, à chaque nœud de chacune de ces deux hiérarchies, est associée une représentation factorielle issue d'une analyse des correspondances particulière (AFC) permettant de visualiser et d'interpréter la scission de ce nœud en ses deux successeurs. Ce couplage entre nœuds et représentations factorielles permet une synthèse plus rapide des résultats.

Une classification croisée du tableau est ainsi obtenue par la construction de ces deux classifications hiérarchiques optimisées et élaguées, édifiées respectivement sur les lignes et les colonnes du tableau. Cette approche se distingue, cependant, des techniques de classifications croisées proposées par Govaert (1984) qui recherchent simultanément une partition des lignes et des colonnes par des méthodes de type nuées dynamiques (Diday, 1971).

Pour chacune des deux hiérarchies obtenues, chaque nœud représente en fait un dipôle composé de deux classes de modalités. Ainsi, en appelant I et J les deux ensembles de modalités définissant les lignes et les colonnes du tableau de contingence, un nœud de la hiérarchie sur I est un dipole composé de deux classes de modalités de I qui ont des associations contraires avec les modalités de J. Loin de compliquer les résultats, cette approche permet également d'obtenir une vue plus synthétique des correspondances entre I et J.

Les hiérarchies optimisées sur I et sur J sont obtenues indépendamment l'une de l'autre par une même méthodologie déjà décrite dans le cadre de la classification factorielle optimisée d'un tableau de mesures (Denimal, 2007). Chaque nœud de la hiérarchie est issu d'une analyse en composantes principales (ACP) particulière. Dans le cadre du traitement d'un tableau de contingence, la méthodologie est cependant adaptée et cette ACP devient équivalente à une analyse factorielle des correspondances particulière. Par contre, les étapes d'élagages des hiérarchies optimisées et d'interprétation de leurs nœuds sont réalisées différemment, à partir de tests conditionnels exacts basés sur le modèle hypergéométrique.

Le calcul des p-valeurs associées à ces tests est réalisé, dans cet article, de manière approchée à partir d'un échantillon de tableaux de contingence de marges fixées. Cet échantillon est obtenu à partir de l'algorithme de Patefield (1981). Cet algorithme est rapide et donne également la probabilité du tableau extrait. On sait par ailleurs que le nombre de tableaux de contingence à marges fixées tend rapidement à devenir trés élevé rendant le calcul exact de cette p-valeur infaisable pour des marges élevées (Mitcell Gail et Nathan Mantel, 1977). En ce qui concerne le calcul de cette p-valeur associée à un test exact de Fisher généralisé aux tableaux de contingence à r lignes et c colonnes, il faut citer l'algorithme proposé par Mehta et Patel (1983) basé

sur une représentation en réseau de l'ensemble des tableaux concernés. Cet algorithme permet non seulement un calcul exact plus rapide de cette p-valeur, mais rend ce calcul faisable dans certains cas où d'autres méthodes le déclarent impossible.

La méthodologie proposée se décompose en plusieurs étapes :

- Construction des hiérarchies dites initiales sur les ensembles I et J du tableau de contingence k_{IJ}.
- Optimisation de ces deux hiérarchies
- Élagages mutuels de ces deux hiérarchies
- Interprétations mutuelles des nœuds et classes des deux hiérarchies.

La méthodologie est illustrée ensuite par un exemple. Le tableau de contingence choisi est celui des votes des 96 départements français pour les différents candidats aux élections présidentielles de 1995.

2. Hiérarchie initiale sur J et optimisation

La construction des hiérarchies initiales H_I et H_J et leur optimisation sont obtenues par la même méthode. Cette dernière sera présentée, dans ce paragraphe, dans le cadre de la hiérarchie sur J.

2.1. Définitions préliminaires

Le tableau de contingence croisant les ensembles I et J est noté k_{IJ}.

Ses effectifs marginaux sur I, son effectif total et ses fréquences marginales seront notés :

$\forall i \in I,\ k(i) = \sum_{j \in J} k(i,j)$.

$k = \sum_{i \in I} k(i)$

$\forall i \in I,\ f_i = \frac{k(i)}{k}$

Ces notations seront conservées dans la totalité de l'article.

La construction de la hiérarchie initiale sur J ne se fera pas directement sur k_{IJ}, mais sur un ensemble de tableaux de la forme $\left\{K\left[j,\overline{j}\right] / j \in J\right\}$, où $K\left[j,\overline{j}\right]$ croisant I et $\left\{j,\overline{j}\right\}$ se déduit comme suit de k_{IJ} : $\forall i \in I$,

$K(i,j) = k(i,j)$

$K(i,\overline{j}) = \sum_{\substack{j' \in J \\ j' \neq j}} k(i,j')$

La colonne $\overline{j}$ est la colonne cumulant les colonnes $j\prime$ de k_{IJ} différentes de j.

Nous introduirons ci-dessous des tableaux $K\left[q,\overline{q}\right]$ plus généraux, mais définis de manière analogue. Les propriétés de ces tableaux sont d'abord explicitées.

2.2. Définition et propriétés des tableaux $K[q, \overline{q}]$

2.2.1. Définitions et notations 1

a) $K[q, \overline{q}]$ est un tableau croisant les ensembles I et $\{q, \overline{q}\}$ tel que :
$\forall i \in I,\ K(i,q) + K(i,\overline{q}) = k(i)$.
Autrement dit, les tableaux $K[q, \overline{q}]$ et k_{IJ} ont les mêmes effectifs marginaux sur I.
b) On pose : $K(q) = \sum_{i \in I} K(i,q)$, $K(\overline{q}) = \sum_{i \in I} K(i,\overline{q})$

2.2.2. Propriété et définition des variables y^q et $y^{\overline{q}}$

L'analyse des correspondances (AFC) du tableau $K[q, \overline{q}]$ *gènère un unique facteur sur I non trivial défini au signe prés, par :*

$$\forall i \in I,\ y^q(i) = \frac{\sqrt{K(q).K(\overline{q})}}{k(i)} \cdot \left[\frac{K(i,q)}{K(q)} - \frac{K(i,\overline{q})}{K(\overline{q})} \right]$$

On vérifie que $y^{\overline{q}}(i) = -y^q(i)$

Démonstration. — L'unique vecteur axial non trivial issu de l'AFC de $K[q, \overline{q}]$ est le vecteur u, normé au sens de la métrique du chi-deux, et orthogonal à $\begin{pmatrix} K(q) \\ K(\overline{q}) \end{pmatrix}$:

$$u = \begin{pmatrix} u_q \\ u_{\overline{q}} \end{pmatrix} = \frac{\sqrt{K(q).K(\overline{q})}}{k} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

La coordonnée $y^q(i)$ de la ligne i est égale au produit scalaire au sens du chi-deux entre u et $\begin{pmatrix} \frac{K(i,q)}{k(i)} \\ \frac{K(i,\overline{q})}{k(i)} \end{pmatrix}$.

Autrement dit, $y^q(i) = \frac{k}{K(q)}.u_q.\frac{K(i,q)}{k(i)} + \frac{k}{K(\overline{q})}.u_{\overline{q}}.\frac{K(i,\overline{q})}{k(i)}$

Ce qui donne finalement : $y^q(i) = \frac{\sqrt{K(q).K(\overline{q})}}{k(i)} \cdot \left[\frac{K(i,q)}{K(q)} - \frac{K(i,\overline{q})}{K(\overline{q})} \right]$

2.2.3. Propriétés

Propriété 1. — *La variable y^q se définit encore comme suit : $\forall i \in I$,*

$$y^q(i) = \sqrt{\frac{K(q)}{K(\overline{q})}} \cdot \left[\frac{K(i,q)}{f_i.K(q)} - 1 \right]$$

Démonstration. —

$$\begin{aligned} \frac{K(i,q)}{K(q)} - \frac{K(i,\overline{q})}{K(\overline{q})} &= \frac{K(i,q)}{K(q)} - \frac{k(i) - K(i,q)}{K(\overline{q})} \\ &= K(i,q).\left[\frac{1}{K(q)} + \frac{1}{K(\overline{q})} \right] - \frac{k(i)}{K(\overline{q})} \end{aligned}$$

Ce qui vaut encore :

$$\frac{k.K(i,q)}{K(q).K(\overline{q})} - \frac{k(i)}{K(\overline{q})} = \frac{k(i)}{K(\overline{q})}.\left[\frac{k.K(i,q)}{K(q).k(i)} - 1\right] = \frac{k(i)}{K(\overline{q})}.\left[\frac{K(i,q)}{K(q).f_i} - 1\right]$$

En remplaçant $\frac{K(i,q)}{K(q)} - \frac{K(i,\overline{q})}{K(\overline{q})}$ par l'expression trouvée ci-dessus, dans la formule définissant $y^q(i)$, on obtient la propriété 1.

PROPRIÉTÉ 2. — *Le tableau $K[q,\overline{q}]$ se définit à l'aide de y^q , $K(q)$, $K(\overline{q})$ et des fréquences f_i par les formules : $\forall i \in I$,*

$$K(i,q) = f_i.K(q).\left[1 + \sqrt{\frac{K(\overline{q})}{K(q)}}.y^q(i)\right]$$

$$K(i,\overline{q}) = f_i.K(\overline{q}).\left[1 - \sqrt{\frac{K(q)}{K(\overline{q})}}.y^q(i)\right]$$

Démonstration. —

a) Calculons d'abord les coordonnées $G(q)$ et $G(\overline{q})$ des colonnes q et $\overline{q}$ sur l'unique axe factoriel issu de l'AFC de $K[q,\overline{q}]$.

Ces coordonnées vérifient le système suivant où λ est l'inertie du tableau $K[q,\overline{q}]$:

$$\begin{cases} K(q).G(q) + K(\overline{q}).G(\overline{q}) = 0 \\ \frac{K(q)}{k}.G^2(q) + \frac{K(\overline{q})}{k}.G^2(\overline{q}) = \lambda \end{cases}$$

La résolution de ce système donne : $G(q) = \sqrt{\frac{K(\overline{q})}{K(q)}}.\lambda$; $G(\overline{q}) = -\sqrt{\frac{K(q)}{K(\overline{q})}}.\lambda$

b) Les formules demandées ne sont alors que l'application de la formule de reconstitution de l'AFC appliquée à $K[q,\overline{q}]$.

2.3. Compromis $K[q_0,\overline{q_0}]$ de deux tableaux $K[q_1,\overline{q_1}]$ et $K[q_2,\overline{q_2}]$

2.3.1. Le tableau $K[q_1,q_2,\overline{q_1},\overline{q_2}]$

$K[q1,q2,\overline{q1},\overline{q2}]$ est un tableau croisant les ensembles I et $\{q1,q2,\overline{q1},\overline{q2}\}$ juxtaposant deux tableaux $K[q_1,\overline{q_1}]$ et $K[q_2,\overline{q_2}]$.

$$K[q1,q2,\overline{q1},\overline{q2}] = I \left\{ \begin{array}{|c|c|} \hline K[q_1,\overline{q_1}] & K[q_2,\overline{q_2}] \\ \hline \end{array} \right.$$

En conséquence, en conservant la même notation $K(i,q)$ pour un élément du tableau $K[q_1,q_2,\overline{q_1},\overline{q_2}]$, on déduit :

$$\forall i \in I,\ K(i,q_1) + K(i,\overline{q_1}) = K(i,q_2) + K(i,\overline{q_2}) = k(i).$$

$K[q1,q2,\overline{q1},\overline{q2}]$ sera soumis à l'analyse des correspondances. $\forall q \in \{q_1,q_2\}$, chaque couple de points $(q,\overline{q})$ du nuage des colonnes est un dipole constitué de deux points alignés avec le centre de gravité de ce nuage.

2.3.2. Le tableau $Y\,[q,q']$, $q \in \{q_1, \overline{q_1}\}$ et $q' \in \{q_2, \overline{q_2}\}$

Comme $y^{\overline{q_1}} = -y^{q_1}$ et $y^{\overline{q_2}} = -y^{q_2}$, il est toujours possible de choisir $q \in \{q_1, \overline{q_1}\}$et $q\prime \in \{q_2, \overline{q_2}\}$ de façon à ce que la covariance $cov(y^q, y^{q\prime}) = \sum_{i\in I} f_i.y^q(i).y^{q'}(i)$ soit positive. Le tableau $Y\,[q,q']$ est le tableau croisant I et $\{q,q'\}$ constitué des deux variables y^q et $y^{q\prime}$ pour lesquelles $cov(y^q, y^{q\prime}) \geqslant 0$.
On soumet $Y\,[q,q']$ à l'analyse en composantes principales non normée (ACP), chaque élément $i \in I$ étant muni du poids f_i et la métrique dans R^2 étant la métrique euclidienne classique. Dans ce cadre, le facteur sur I associé à la plus grande valeur propre s'écrit sous la forme :

$$y^{q_0} = \alpha.y^q + \beta.y^{q'} \text{ avec } \alpha \geqslant 0,\ \beta \geqslant 0 \text{ et } \alpha^2 + \beta^2 = 1$$

2.3.3. Propriété 3

L'analyse des correspondances de $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$ et l'analyse en composantes principales non normée de $\frac{1}{\sqrt{2}}.Y[q,q']$ sont équivalentes. Elles génèrent les mêmes valeurs propres et les mêmes facteurs sur I.

Démonstration. — Il suffit de démontrer que les deux nuages sur I issus de ces deux tableaux ont même triple (I, f_I, d_{II}).
Il est clair que pour ces deux tableaux les poids f_i des éléments de I sont identiques.
Calculons dans les deux cas les distances carrées $d^2\,(i,i')$:
a) Dans le cadre de l'analyse en composantes principales de $\frac{1}{\sqrt{2}}Y\,(q,q')$,

$$d^2\,(i,i') = \frac{1}{2}[y^q(i) - y^q(i')]^2 + \frac{1}{2}[y^{q'}(i) - y^{q'}(i')]^2$$

comme $y^q(i) = \varepsilon.y^{q_1}(i)$ et $y^{q'}(i) = \varepsilon.y^{q_2}(i)$, $\forall i \in I$ avec $\varepsilon = \pm 1$, on peut écrire que :

$$d^2\,(i,i') = \frac{1}{2}\,[y^{q_1}(i) - y^{q_1}(i')]^2 + \frac{1}{2}\,[y^{q_2}(i) - y^{q_2}(i')]^2$$

En remplaçant y^q par sa définition, on a :

$$d^2\,(i,i') = \frac{K(q_1).K(\overline{q_1})}{2}.\left[\frac{K(i,q_1)}{k(i).K(q_1)} - \frac{K(i,\overline{q_1})}{k(i).K(\overline{q_1})} - \frac{K(i',q_1)}{k(i').K(q_1)} + \frac{K(i',\overline{q_1})}{k(i').K(\overline{q_1})}\right]^2$$

$$+\frac{K(q_2).K(\overline{q_2})}{2}\left[\frac{K(i,q_2)}{k(i).K(q_2)} - \frac{K(i,\overline{q2})}{k(i).K(\overline{q_2})} - \frac{K(i',q_2)}{k(i').K(q_2)} + \frac{K(i',\overline{q_2})}{k(i').K(\overline{q_2})}\right]^2$$

En appliquant la définition de $K(i,\overline{q})$, on peut écrire :

$\forall i \in I,\ \forall q \in \{q_1, q_2\}$,

$$\frac{K(i,\overline{q})}{k(i).K(\overline{q})} = \frac{k\,(i)}{k(i).K(\overline{q})} - \frac{K(i,q)}{k(i).K(\overline{q})} = \frac{1}{K(\overline{q})} - \frac{K(i,q)}{k(i).K(\overline{q})}$$

$$\frac{K(i,q)}{k(i).K(q)} - \frac{K(i,\overline{q})}{k(i).K(\overline{q})} = \frac{K(i,q)}{k(i)}.\left[\frac{1}{K(q)} + \frac{1}{K(\overline{q})}\right] - \frac{1}{K(\overline{q})}$$

$$\frac{K(i,q)}{k(i).K(q)} - \frac{K(i,\overline{q})}{k(i).K(\overline{q})} - \frac{K(i',q)}{k(i').K(q)} + \frac{K(i',\overline{q})}{k(i').K(\overline{q})}$$
$$= \left[\frac{1}{K(q)} + \frac{1}{K(\overline{q})}\right] . \left[\frac{K\,(i,q)}{k(i)} - \frac{K(i',q)}{k(i')}\right]$$

En remarquant que $K(q) + K(\overline{q}) = k$, on déduit :

$$\frac{K(i,q)}{k(i).K(q)} - \frac{K(i,\overline{q})}{k(i).K(\overline{q})} - \frac{K(i',q)}{k(i').K(q)} + \frac{K(i',\overline{q})}{k(i').K(\overline{q})}$$
$$= \frac{k}{K(q).K(\overline{q})} . \left[\frac{K\,(i,q)}{k(i)} - \frac{K(i',q)}{k(i')}\right]$$

Finalement, on obtient :

$$d^2\,(i,i') = \sum_{q\in\{q_1,q_2\}} \frac{k^2}{2.K(q).K(\overline{q})} . \left[\frac{K(i,q)}{k(i)} - \frac{K(i',q)}{k(i')}\right]^2$$

b) Dans le cadre de l'analyse des correspondances de $K\,[q_1, q_2, \overline{q_1}, \overline{q_2}]$, la distance carrée au sens du chi-deux s'écrit (puisque la fréquence de q vaut alors $\frac{K(q)}{2.k}$) :

$$d^2\,(i,i') = \sum_{q\in\{q_1,q_2,\overline{q_1},\overline{q_2}\}} \frac{2k}{K(q)} . \left[\frac{K(i,q)}{2.k(i)} - \frac{K(i',q)}{2.k(i')}\right]^2$$

En utilisant la définition de $K(i,\overline{q})$,on peut écrire que :

$$\frac{K(i,\overline{q})}{k(i)} - \frac{K(i',\overline{q})}{k(i')} = \frac{K(i',q)}{k(i')} - \frac{K(i,q)}{k(i)}, \forall q \in \{q_1, q_2\}$$

Autrement dit, on déduit :

$$d^2\,(i,i') = \sum_{q\in\{q_1,q_2\}} \left(\frac{2k}{K(q)} + \frac{2k}{K(\overline{q})}\right) . \left[\frac{K(i,q)}{2.k(i)} - \frac{K(i',q)}{2.k(i')}\right]^2$$

Comme $K(q) + K(\overline{q}) = k$, on trouve finalement :

$$d^2\,(i,i') = \sum_{q\in\{q_1,q_2\}} \frac{k^2}{2.K(q).K(\overline{q})} . \left[\frac{K(i,q)}{k(i)} - \frac{K(i',q)}{k(i')}\right]^2$$

Remarque. — La propriété 3 est un résultat classique observé par exemple dans le cas du dedoublement d'un tableau de notes ou dans le cas d'un tableau disjonctif complet dont les questions n'ont que deux modalités (Benzecri, 1976 ; Lebart, Morineau, Piron 1995)

2.3.4. Représentations graphiques associées

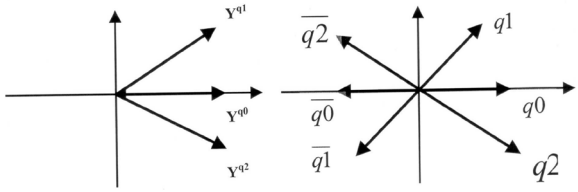


FIG 1. — Représentations factorielles associées aux deux analyses équivalentes.

Les modalités $q \in \{q_1, \overline{q_1}\}$ *et* $q' \in \{q_2, \overline{q_2}\}$ sont choisies telles que $cov(y^q, y^{q'}) \geqslant 0$. La figure 1 ci-dessous présente le cas particulier $q = q_1$ et $q' = q_2$.

La définition des modalités q_0 et $\overline{q_0}$ et leurs positions sur l'axe 1 seront explicitées ci-dessous.

Le lemme suivant n'est qu'un intermédiaire technique nécessaire à la démonstration de la conséquence 2.4 e).

LEMME. — *On considère* $q \in \{q_1, \overline{q_1}\}$ *et* $q' \in \{q_2, \overline{q_2}\}$ *choisies tels que* $cov(y^q, y^{q'}) \geqslant 0$. *Leurs coordonnées* $G(q), G(q')$ *sur le premier axe factoriel issu de l'AFC de* $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$ *valent :*

$$G(q) = \sqrt{\frac{K(\overline{q})}{K(q)}}.\sqrt{2.\lambda_1}\alpha\,;\; G(q') = \sqrt{\frac{K(\overline{q'})}{K(q')}}.\sqrt{2.\lambda_1}\beta$$

où λ_1 est la plus grande valeur propre issue de l'ACP de $\frac{1}{\sqrt{2}}.Y[q, q']$ (ou de l'AFC de $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$).

De même, $G(\overline{q}) = -\sqrt{\frac{K(q)}{K(\overline{q})}}.\sqrt{2.\lambda_1}\alpha\,;\; G(\overline{q'}) = -\sqrt{\frac{K(q')}{K(\overline{q'})}}.\sqrt{2.\lambda_1}\beta$

où $\overline{q}$ *et* $\overline{q\prime}$ *sont les éléments tels que* $\{q, \overline{q}\} = \{q_1, \overline{q_1}\}$ *et* $\{q', \overline{q'}\} = \{q_2, \overline{q_2}\}$

Démonstration. — On sait que l'ACP de $\frac{1}{\sqrt{2}}Y[q, q\prime]$ génère un facteur sur I, noté $\frac{1}{\sqrt{2}}.y^{q0}$, associé à la plus grande valeur propre λ_1, tel que $y^{q0} = \alpha.y^q + \beta y^{q'}$ avec $\alpha \geqslant 0$, $\beta \geqslant 0$ et $\alpha^2 + \beta^2 = 1$.

Les propriétés classiques de l'ACP montrent que les coordonnées sur le premier axe factoriel des colonnes q et $q\prime$ issu de l'ACP de $\frac{1}{\sqrt{2}}Y[q, q\prime]$ sont respectivement $\sqrt{\lambda_1}.\alpha$ et $\sqrt{\lambda_1}.\beta$ où λ_1 est la première valeur propre issue de cette analyse.

La coordonnée de q s'obtient également par la formule :

$$\sqrt{\lambda_1}.\alpha = \frac{1}{2.\sqrt{\lambda_1}} \sum_{i \in I} f_i.y^q(i).y^{q_0}(i)$$

D'autre part, la variable y^q s'écrit encore sous la forme (propriété 1) :

$$y^q(i) = \sqrt{\frac{K(q)}{K(\overline{q})}}.\left[\frac{K(i,q)}{f_i.K(q)} - 1\right]$$

En conséquence, en remplaçant $y^q(i)$ par sa valeur rappelée ci-dessus, et en utilisant le fait que le facteur y^{q_0} est centrée (autrement dit, $\sum_{i\in I} f_i.y^{q_0}(i) = 0$), la coordonnée de q sur le premier axe factoriel issu de l'ACP de $\frac{1}{\sqrt{2}}Y[q,q\prime]$ devient :

$$\sqrt{\lambda_1}.\alpha = \sqrt{\frac{K(q)}{2.K(\overline{q})}}.\left(\frac{1}{\sqrt{\lambda_1}}\sum_{i\in I}\frac{K(i,q)}{K(q)}.\frac{y^{q_0}(i)}{\sqrt{2}}\right)$$

Or, dans le cadre de l'AFC de $K[q1,q2,\overline{q1},\overline{q2}]$, $\frac{y^{q_0}}{\sqrt{2}}$ reste le premier facteur sur I (Propriété 3). La coordonnée $G(q)$ de q s'obtient à partir de la formule de transition : $G(q) = \frac{1}{\sqrt{\lambda_1}}\sum_{i\in I}\frac{K(i,q_1)}{K(q_1)}.\frac{y^{q_0}(i)}{\sqrt{2}}$

On en déduit :

$$\sqrt{\frac{K(q)}{2.K(\overline{q})}}G(q) = \sqrt{\lambda_1}\alpha.$$

Ce qui donne la formule cherchée : $G(q) = \sqrt{\frac{K(\overline{q})}{K(q)}}.\sqrt{2.\lambda_1}\alpha$.

Un raisonnement analogue conduit aux formules exprimant $G(q')$, $.G(\overline{q})$, $G(\overline{q'})$.

2.3.5. Définition du $K[q_0,\overline{q_0}]$, tableau compromis de $K[q_1,\overline{q_1}]$ et $K[q_2,\overline{q_2}]$

Le tableau compromis de $K[q_1,\overline{q_1}]$ et $K[q_2,\overline{q_2}]$ est le tableau de la forme $K[q_0,\overline{q_0}]$ vérifiant :

1) $\forall i \in I$, $K(i,q_0)+K(i,\overline{q_0}) = k(i)$, où $k(i)$ est le total de la ligne i du tableau initial k_{IJ}.

2) Le facteur sur I, non trivial, issu de l'analyse des correspondances de $K[q_0,\overline{q_0}]$ est y^{q_0}, facteur sur I associé à la plus grande valeur propre, issu de l'analyse en composantes principales non normée de $Y[q,q']$ où $q \in \{q_1,\overline{q_1}\}$ et $q' \in \{q_2,\overline{q_2}\}$ sont choisis de façon à ce que la covariance $cov(y^q,y^{q'})$ soit positive ou nulle.

3) Les totaux $K(q_0)$ et $K(\overline{q_0})$ des deux colonnes de $K[q_0,\overline{q_0}]$ valent :
$K(q_0) = \alpha^2.K(q) + \beta^2.K(q')$
$K(\overline{q_0}) = k - K(q_0)$
où k est le total général du tableau initial k_{IJ}, $K(q)$ et $K(q\prime)$ les totaux des colonnes q et $q\prime$ des tableaux $K[q_1,\overline{q_1}]$ et $K[q_2,\overline{q_2}]$, α et β les coefficients positifs ou nuls tels que $y^{q_0} = \alpha.y^q + \beta.y^{q'}$ et $\alpha^2+\beta^2 = 1$.

2.4. Conséquences et interprétations

a) Le tableau compromis $K\left[q_0, \overline{q_0}\right]$ peut se définir à partir des quantités $f_i, K(q_0), K(\overline{q_0})$ et y^{q_0} par la formule de reconstitution classique issue de l'analyse des correspondances de $K\left[q_0, \overline{q_0}\right]$ (voir propriété 1) : $\forall i \in I$,

$$K\left(i, q_0\right) = f_i.K(q_0).\left[1 + \sqrt{\frac{K\left(\overline{q_0}\right)}{K\left(q_0\right)}}.y^{q_0}\left(i\right)\right]$$

$$K\left(i, \overline{q_0}\right) = f_i.K(\overline{q_0}).\left[1 - \sqrt{\frac{K\left(q_0\right)}{K\left(\overline{q_0}\right)}}.y^{q_0}\left(i\right)\right]$$

b) L'unique facteur sur I, y^{q0}, issu de l'analyse des correspondances du compromis $K\left[q_0, \overline{q_0}\right]$ résulte de l'analyse en composantes principales des deux facteurs y^{q1}et y^{q2} issus respectivement des analyses des correspondances des tableaux $K\left[q_1, \overline{q_1}\right]$ et $K\left[q_2, \overline{q_2}\right]$ ce qui justifie donc son nom de compromis.

c) Les deux coefficients positifs ou nuls α et β représentent la contribution de chacune des variables y^qet $y^{q'}$ à la construction du compromis y^{q0}. En effet, les formules classiques de l'ACP montrent que :

$$\alpha = \frac{cov(y^{q0}, y^q)}{\sqrt{cov^2(y^{q0}, y^q) + cov^2(y^{q0}, y^{q'})}}$$

$$\beta = \frac{cov(y^{q0}, y^{q'})}{\sqrt{cov^2(y^{q0}, y^q) + cov^2(y^{q0}, y^{q'})}}$$

d) De même, le poids $\frac{K(q_0)}{k}$ associé à la colonne q_0 de $K\left[q_0, \overline{q_0}\right]$ est la moyenne pondérée des poids $\frac{K(q)}{k}$ et $\frac{K(q')}{k}$ des colonnes q et q', suivant les coefficients α^2 et β^2 de somme 1. En conséquence, le poids de la colonne compromis q_0 s'interpréte aussi comme le compromis des poids de q et de $q\prime$.

D'autre part, si l'on suppose, par exemple, que $q = q_1$et $q\prime = q_2$, on vérifie que :

$K(\overline{q_0}) = k - K(q_0) = k.(\alpha^2 + \beta^2) - (\alpha^2.K(q_1) + \beta^2.K(q2)) = \alpha^2.K(\overline{q_1}) + \beta^2.K(\overline{q_2})$

$K(\overline{q_0})$ s'interpréte aussi comme le compromis des quantités $K(\overline{q_1})$ et $K(\overline{q_2})$.

e) Dans le cadre de l'analyse des correspondances du tableau $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$, plaçons le tableau $K\left[q_0, \overline{q_0}\right]$ en colonnes supplémentaires. on montre ci-dessous que les points q_0 et $\overline{q_0}$ représentant les profils des colonnes de $K\left[q_0, \overline{q_0}\right]$ se positionnent sur le premier axe factoriel issu de l'AFC de $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$ de part et d'autre de l'origine (voir Figure 1). Ainsi, le dipôle $(q_0, \overline{q_0})$ s'interpréte bien comme le compromis des dipôles $(q, \overline{q})$ et $(q', \overline{q'})$.

En effet, on déduit facilement de la propriété a) précédente :

$$f_I^{q_0} - f_I = \sqrt{\frac{K(\overline{q_0})}{K\left(q_0\right)}}.\left(f_i.y^{q_0}\left(i\right)\right)_{i \in I}$$

$$f_I^{\overline{q_0}} - f_I = -\sqrt{\frac{K(q_0)}{K\left(\overline{q_0}\right)}}.\left(f_i.y^{q_0}\left(i\right)\right)_{i \in I}$$

où $f_I^{q_0}$et $f_I^{\overline{q_0}}$ représentent les profils de q_0 et $\overline{q_0}$ et où f_I est le vecteur des fréquences f_i.

D'autre part, y^{q_0} est le premier facteur sur I issu de l'ACP de $Y\,[q, q']$. Donc, $\frac{1}{\sqrt{2}}.y^{q_0}$ est le premier facteur sur I issu de l'ACP de $\frac{1}{\sqrt{2}}.Y\,[q, q']$, et par conséquent celui issu également de l'AFC de $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$ (propriété 3).

Considérant cette dernière analyse, si l'on note λ_1 la plus grande valeur propre obtenue, le vecteur axial qui lui est associé s'écrit : $u_I = \left(\frac{f_i.y^{q_0}(i)}{\sqrt{2.\lambda_1}}\right)_{i\in I}$

Des égalités précédentes définissant $f_I^{q_0} - f_I$ et $f_I^{\overline{q_0}} - f_I$, on déduit :

$$f_I^{q_0} - f_I = \sqrt{\frac{K(\overline{q_0})}{K\,(q_0)}}.\sqrt{2.\lambda_1}.u_I$$

$$f_I^{\overline{q_0}} - f_I = -\sqrt{\frac{K(q_0)}{K\,(\overline{q_0})}}.\sqrt{2.\lambda_1}.u_I$$

q_0 et $\overline{q_0}$ se positionnent donc sur le premier axe factoriel et admettent les coordonnées :

$$G(q_0) = \sqrt{2.\lambda_1}.\sqrt{\frac{K(\overline{q_0})}{K\,(q_0)}}$$

$$G(\overline{q_0}) = -\sqrt{2.\lambda_1}.\sqrt{\frac{K(q_0)}{K\,(\overline{q_0})}}$$

f) Nous allons à présent montrer que la modalité q_0 (respectivement $\overline{q_0}$) peut s'interpréter comme le compromis des modalités q et q' (respectivement $\overline{q}$ et $\overline{q'}$), $q \in \{q_1, \overline{q_1}\}$ *et* $q' \in \{q_2, \overline{q_2}\}$ (voir §2.3.4)

Introduisons les fréquences $f_{q_0} = \frac{K(q_0)}{k}$, $f_q = \frac{K(q)}{k}$, $f_{q'} = \frac{K(q')}{k}$ et notons $G(q)$ et $G(q\prime)$ les coordonnées de q et de q' sur le premier axe issu de l'AFC de $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$.

On montre alors que :

$$f_{q_0}.\left\|f_I^{q_0} - f_I\right\|^2 = f_{q_0}.\left[G(q_0)\right]^2 = f_q.\left[G(q)\right]^2 + f_{q'}.\left[G(q')\right]^2$$

$$f_{\overline{q_0}}.\left\|f_I^{\overline{q_0}} - f_I\right\|^2 = f_{\overline{q_0}}.\left[G(\overline{q_0})\right]^2 = f_{\overline{q}}.\left[G(\overline{q})\right]^2 + f_{\overline{q'}}.\left[G(\overline{q'})\right]^2$$

où $\overline{q}$ et $\overline{q\prime}$ sont les éléments tels que $\{q, \overline{q}\} = \{q_1, \overline{q_1}\}$ et $\left\{q', \overline{q'}\right\} = \{q_2, \overline{q_2}\}$.

En utilisant le lemme précédent et les formules données ci-dessus exprimant $G(q_0)$ et $G(\overline{q_0})$, on montre facilement que les égalités précédentes s'écrivent encore sous la forme :

$$K(q_0) = \alpha^2.K(q) + \beta^2.K\,(q')$$

$$K(\overline{q_0}) = \alpha^2.K(\overline{q}) + \beta^2.K(\overline{q'})$$

Ce qui est la définition des poids attribués à q_0 et $\overline{q_0}$ (voir définition 2.3.5)

g) L'inertie du dipôle compromis $(q_0, \overline{q_0})$ est égale à $2.\lambda_1$ où λ_1 est la première valeur propre issue de l'AFC de $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$.

En effet, on sait que :

$$2.\lambda_1 = f_q.\left[G(q)\right]^2 + f_{q'}.\left[G(q')\right]^2 + f_{\overline{q}}.\left[G(\overline{q})\right]^2 + f_{\overline{q'}}.\left[G(\overline{q'})\right]^2$$

La présence du coefficient 2 s'explique par le fait que les poids de $q_1, q_2, \overline{q_1}, \overline{q_2}$ dans le cadre de l'AFC de $K[q_1, q_2, \overline{q_1}, \overline{q_2}]$ sont respectivement

$$\frac{K(q_1)}{2k}, \frac{K(q_2)}{2k}, \frac{K(\overline{q_1})}{2k}, \frac{K(\overline{q_2})}{2k}.$$

À partir des propriétés f), on déduit que

$$2.\lambda_1 = f_{q_0}.f_{\overline{q_0}}.\left\|f_I^{q_0} - f_I^{\overline{q_0}}\right\|^2 = f_{q_0}.\left\|f_I^{q_0} - f_I\right\|^2 + f_{\overline{q_0}}.\left\|f_I^{\overline{q_0}} - f_I\right\|^2$$

2.5. Algorithme de construction de la hiérarchie initiale sur J

Cet algorithme n'est autre que l'algorithme de classification de variables déja présenté dans un article précédent (Denimal, 2001). Il va se distinguer cependant par le fait que les variables concernées sont ici très particulières.

L'objectif est de construire une classification ascendante hiérarchique sur l'ensemble J du tableau de contingence k_{IJ}. Un ensemble de $card(J)$ tableaux $\left\{K[j,\overline{j}] \ / \ j \in J\right\}$, est d'abord défini à partir du tableau de contingence initial k_{IJ}.

Chaque tableau $K[j,\overline{j}]$ croise les ensembles I et $\left\{j,\overline{j}\right\}$ et se définit à partir du tableau initial k_{IJ} par :

$$\forall i \in I,\ K\left(i,j\right) = k(i,j) \text{ et } K\left(i,\overline{j}\right) = k(i) - k(i,j) \text{ où } k(i) = \sum_{j\in J} k(i,j).$$

La classification proposée présentera plusieurs interprétations. Elle peut être considérée comme la classification des tableaux $K[j,\overline{j}]$, ou encore comme celle des dipoles $(j,\overline{j})$.

À chaque tableau $K[j,\overline{j}]$, sera associée une variable y^j représentant le facteur non trivial sur I issu de l'AFC de $K[j,\overline{j}]$. La définition de y^j, donnée précédemment de manière générale (voir §2.2.2), se transcrit ici de la façon suivante :

$$\forall i \in I,\ y_i^j = \frac{\sqrt{f_j.f_{\overline{j}}}}{f_i}.\left[\frac{K(i,j)}{K(j)} - \frac{K(i,\overline{j})}{K(\overline{j})}\right].$$

avec $K\left(j\right) = \sum_{i\in I} K\left(i,j\right)$, $K\left(\overline{j}\right) = \sum_{i\in I} K(i,\overline{j})$, $k(i) = \sum_{j\in J} k(i,j)$, $k = \sum_{i\in I} k\left(i\right)$,

$f_{i=}\frac{k\left(i\right)}{k}$, $f_j = \frac{K\left(j\right)}{k}$, $f_{\overline{j}} = \frac{K\left(\overline{j}\right)}{k}$

La classification proposée sera également la classification hiérarchique de ces $card(J)$ variables y^j, $j \in J$, suivant l'algorithme déja explicité dans l'article (Denimal, 2001). Cet algorithme basé sur les analyses en composantes principales des tableaux $Y[j_1, j_2]$ regroupant deux variables y^{j_1} et y^{j_2}

présente ici une nouvelle interprétation puisque l'analyse en composantes principales de $Y[j_1, j_2]$ est équivalente à l'analyse des correspondances du tableau $K\left[j_1, j_2, \overline{j_1}, \overline{j_2}\right]$ juxtaposant les tableaux $K[j_1, \overline{j_1}]$ et $K[j_2, \overline{j_2}]$ (Propriété 3).

Chaque variable représentative de classe sera obtenue à partir d'un vecteur $a = (a_j)_{j \in J}$ de coefficients obtenu de manière itérative par l'algorithme. En effet, chaque classe q obtenue $(q \subset J)$ sera représentée par une variable $y^q = \sum_{j \in q} \frac{a_j}{\sqrt{\sum_{j \in q} a_j^2}} . y^j$ ou par un dipole $(q, \overline{q})$ constitué des profils des deux colonnes q et $\overline{q}$ d'un tableau de contingence particulier $K[q, \overline{q}]$. Ce vecteur $a = (a_j)_{j \in J}$ jouera un rôle important dans l'étape d'optimisation de cette hiérarchie.

L'algorithme de construction de la hiérarchie sur l'ensemble J de k_{IJ} se définit comme suit :

Étape 0 : On pose $a^0 = \left(a_j^0\right)_{j \in J}$ telle que $a_j^0 = 1 \ \forall j \in J$.

Étape 1 : - $\forall j \in J$, on considère les tableaux $K\left[j, \overline{j}\right]$ et les variables associées y^j . Pour chaque couple de modalités $(j_1, j_2) \in J \times J$, on réalise l'analyse en composantes principales non normée du tableau $Y[j_1, j_2]$ regroupant les variables y^{j1} et y^{j2}. Cette analyse génère deux valeurs propres $\lambda_1(j1, j2) \geqslant \lambda_2(j1, j2)$ telles que :

$$\lambda_1(j1, j2) = \sum_{i \in I} f_i \left[\alpha_1 y_i^{j1} + \beta_1 y_i^{j2}\right]^2 = var(\alpha_1 y^{j1} + \beta_1 y^{j2}) \text{ avec } \alpha_1^2 + \beta_1^2 = 1.$$

$$\lambda_2(j1, j2) = var(y^{j1}) + var(y^{j2}) - \lambda_1(j1, j2) = var(\beta_1 y^{j1} - \alpha_1 y^{j2})$$

On détermine ensuite le couple (j_1, j_2) pour lequel $\lambda_2(j_1, j_2)$est minimum.

On en déduit alors la série de coefficients a^1.

$$\begin{cases} a^1(j1) = \alpha_1 \\ a^1(j2) = \beta_1 \\ a^1(j) = a^0(j) \ , \ j \neq j1 \ , \ j \neq j2 \end{cases}$$

L'indice du premier nœud $n_1 = (y^{j1}, y^{j2})$ vaut : $\nu(n_1) = \lambda_2(j_1, j_2)$

La variable représentative de ce nœud $n_1$est : $y^q = a_1(j_1).y^{j_1} + a_1(j_2).y^{j_2}$ avec $q = \{j_1, j_2\}$.

D'autre part, l'ACP non normée du tableau $Y[j1, j2]$ est équivalente à l'analyse des correspondances du tableau $K\left[j_1, j_2, \overline{j_1}, \overline{j_2}\right]$ juxtaposant les tableaux $K[j_1, \overline{j_1}]$ et $K[j_2, \overline{j_2}]$ (Propriété 3). La variable représentative y^q est encore le facteur non trivial sur I issu de l'AFC de $K[q, \overline{q}]$ tableau compromis des deux tableaux $K[j_1, \overline{j_1}]$ et $K[j_2, \overline{j_2}]$ (Définition 2.3.5).

Étape k : $(k \in [2, p-1])$. Chaque classe q obtenue aprés $(k-1)$ étapes est représentée par la variable : $y^q = \sum_{j \in q} a_{k-1}(j) . y^j$. Pour chaque couple de variables (y^{q_1}, y^{q_2}), on réalise l'ACP non normée du tableau $Y[q_1, q_2]$ associé. Les deux valeurs propres extraites sont notées : $\lambda_1(q_1, q_2) \geqslant \lambda_2(q_1, q_2)$.

$\lambda_1(q_1,q_2) = \sum_{i\in I} f_i.[\alpha_k.y_i^{q_1} + \beta_k.y_i^{q_2}]^2 = var(\alpha_k.y^{q_1} + \beta_k.y^{q_2})$ avec $\alpha_k^2 + \beta_k^2 = 1$

$\lambda_2(q_1,q_2) = var(y^{q_1}) + var(y^{q_2}) - \lambda_1(q_1,q_2) = var(\beta_k.y^{q_1} - \alpha_k.y^{q_2})$.

On détermine alors le couple $\left(y^{q1}, y^{q2}\right)$ pour lequel $\lambda_2(q_1,q_2)$ est minimum.

On déduit alors :

– une nouvelle série de coefficients $a^k = \left(a_j^k\right)_{j\in J}$ telle que

$$\begin{cases} a_j^k = \alpha_k.a_j^{k-1}, & j \in q1 \\ a_j^k = \beta_k.a_j^{k-1}, & j \in q2 \\ a_j^k = a_j^{k-1}, sinon \end{cases}$$

– l'indice du nouveau nœud formé : $\nu(n_k) = \lambda_2(q_1,q_2)$.

– La variable representative la classe $q_1 \cup q_2$: $y^{q_1\cup q_2} = \sum_{j\in q_1\cup q_2} a_j^k y^j = \alpha_k.y^{q_1} + \beta_k.y^{q_2}$.

De la même manière, l'ACP non normée du tableau $Y[q_1,q_2]$ est équivalente à l'analyse des correspondances du tableau $K[q_1,q_2,\overline{q_1},\overline{q_2}]$ juxtaposant les tableaux $K[q_1,\overline{q_1}]$ et $K[q_2,\overline{q_2}]$ (Propriété 3). La variable représentative y^q avec $q = q_1 \cup q_2$ est encore le facteur non trivial sur I issu de l'AFC de $K[q,\overline{q}]$ tableau compromis des deux tableaux $K[q_1,\overline{q_1}]$ et $K[q_2,\overline{q_2}]$ (Définition 2.3.5).

Par définition, la série a^{p-1} obtenue lors de la dernière étape de l'algorithme sera simplement notée : $a = (a_j)_{j\in J}$. Autrement dit, $\forall j \in J,\ a_{j=}a_j^{p-1}$.

Considérant cet algorithme de construction comme celui de la hiérarchie construite sur les variables y^j , $j \in J$, il vérifie les propriétés suivantes (voir Denimal 2000 ou 2001).

Propriété 4. — *Si on note $\lambda_{1,|J|-1}$ la valeur propre la plus grande issue de l'analyse en composante principales associée au nœud le plus haut ($|J| = card(J)$), on a :*

a) Les indices d'agrégation des nœuds $n_1, n_2, ..., n_{|J|-1}$ (rangés du nœud le plus bas vers le plus haut) forment une suite croissante, majorée par $\lambda_{1,|J|-1}$:

$\nu(n_1) \leqslant \nu(n_2) \leqslant ..... \leqslant \nu\left(n_{|J|-1}\right) \leqslant \lambda_{1,|J|-1}$

b) $\sum_{k=1}^{|J|-1} \nu(n_k) + \lambda_{1,|J|-1} = \sum_{j\in J} var\left(y^j\right)$

c) Les techniques des graphes réductibles (Bruynhooge 1978) ou celle des voisins réciproques (Juan, 1982) peuvent être appliquées pour accélérer la construction de la hiérarchie.

2.6. Optimisation de la hiérarchie initiale construite sur J

Cette optimisation est obtenue en appliquant l'algorithme explicité en détails dans l'article Denimal (2007). Nous le rappelons ci-dessous brièvement.

La phase d'optimisation de la hiérarchie initiale a pour but de recalculer le contenu des classes et le vecteur de coefficients $a = (a_j)_{j\in J}$ de façon à ce

que la variable représentative $z^q = \sum_{j \in q} \frac{a_j}{\sqrt{\sum_{j \in q} (a_j)^2}} . y^j$ de chaque classe q soit de variance maximum. En notant $\mathcal{P}_0$,$\mathcal{P}_1$,.....$\mathcal{P}_{p-1}$ la succession des partitions associées à la hiérarchie et $\nu(k)$ l'indice du nœud n_k, le critère que l'on se propose de maximiser est :

$$Q = \sum_{k=1}^{p-1} \nu(k) . \sum_{q \in \mathcal{P}_k} var(z^q)$$

Le processus d'optimisation recherche une hiérarchie sur J et un vecteur a associé maximisant Q suivant une technique du type des nuées dynamiques (Diday, 1971).

3. Élagage des hiérarchies construites sur I et sur J

Les deux hiérarchies H_I et H_J optimisées sont construites indépendamment l'une de l'autre, de façon analogue, suivant la méthodologie exposée au §2. Cette approche est différente de celle adoptée dans le cadre de la classification factorielle d'un tableau de mesures (Denimal,2007) où la hiérarchie des individus se déduisait de celle des variables. Dans le cas d'un tableau de contingence, nous avons préféré opérer de manière symétrique et appliquer un traitement analogue aux lignes et aux colonnes, suivant ainsi la même démarche que l'analyse des correspondances du tableau.

L'étape suivante consiste à élaguer mutuellement les deux hiérarchies optimisées H_I et H_J de façon à ne conserver que leurs nœuds statistiquement significatifs. Cet élagage repose sur l'utilisation d'un test statistique basé sur une mesure d'association entre deux nœuds l'un de H_I et l'autre de H_J. Ce test et l'indice associé sont explicités en détails dans Denimal (1997) et dans Denimal-Camiz (2001).

Considérons un nœud de l'une des deux hiérarchies H_I ou H_J. Par construction, ce nœud est représenté par un dipôle $(q, \overline{q})$ ou par la variable y^q associée. D'autre part, ce nœud regroupe un sous ensemble c de modalités de I ou de J. $\forall j \in c$, suivant le signe de la covariance $cov(y^j, y^q)$, on répartit les modalités de c en deux classes c_1 et c_2. le dipôle (c_1, c_2) est appelé le dipôle de modalités associé au nœud considéré. Dans la partie « élagage des hiérarchies » (§3) ainsi que dans celle « interprétation des classes » (§4), les dipôles associés aux nœuds seront ces dipôles de modalités.

3.1. Association entre deux classes $p \subset I$ et $q \subset J$

À partir du tableau k_{IJ} initial, on introduit les notations classiques suivantes :

$$\forall i \in I,\ k(i) = \sum_{j \in J} k(i,j)\ ;\ \forall j \in J,\ k(j) = \sum_{i \in I} k(i,j)\ ;\ k = \sum_{i \in I} k(i) = \sum_{j \in J} k(j)$$

$p \subset I,\ k(p) = \sum_{i \in p} k(i);\ q \subset J,\ k(q) = \sum_{j \in q} k(j);\ p \subset I, q \subset J,\ k(p,q) = \sum_{\substack{i \in p \\ j \in q}} k(i,j)$

Étant donné deux classes non vides $p \subset I$ et $q \subset J$, on introduit $A_{pq} = \frac{k.k(p,q)}{k(p).k(q)}$.

Cette quantité s'interpréte comme une mesure d'association entre p et q : A_{pq} est plus grand (respectivement plus petit) que 1 selon que $k(p,q)$ est plus grand (respectivement plus petit) que $\frac{k(p).k(q)}{k}$, effectif théorique moyen dans le cadre du modèle hypergéométrique.

Chaque nœud de H_I ou H_J peut s'interpréter comme un dipole opposant deux classes de modalités de I ou de J. Chacune de ces deux classes peut également s'identifier à l'ensemble des individus vérifiant les modalités de la classe.

Soit $(p, p\prime)$ et $(q, q\prime)$ deux dipoles représentant deux nœuds l'un de H_I et l'autre de H_J :

$$p \subset I,\ p' \subset I,\ p \cap p' = \emptyset \text{ et } q \subset J,\ q' \subset J,\ q \cap q' = \emptyset.$$

Nous noterons $\overline{p \cup p'}$ (respectivement $\overline{q \cup q'}$) l'ensemble des modalités de I n'appartenant pas à $p \cup p'$ (resp. l'ensemble des modalités de J n'appartenant pas à $q \cup q'$).

Quatre cas particuliers peuvent se produire :

Cas n°1 : Les six classes $p, p\prime, \overline{p \cup p'}, q, q', \overline{q \cup q'}$ sont non vides

Cas n°2 : L'une des trois classes $p, p\prime, \overline{p \cup p'}$ est vide et les trois classes $q, q', \overline{q \cup q'}$ sont non vides

Cas n°3 : L'une des trois classes $q, q', \overline{q \cup q'}$ est vide et les trois classes $p, p\prime, \overline{p \cup p'}$ sont non vides

Cas n°4 : L'une des trois classes $p, p\prime, \overline{p \cup p'}$ est vide ainsi que l'une des trois classes $q, q', \overline{q \cup q'}$.

Le cas n°2 survient par exemple lorsque le dipôle $(p, p\prime)$ associé à un nœud de H_I n'est composé que d'une seule classe (autrement dit p ou p' est vide) ou lorsque le nœud de H_I est le sommet de l'arbre (autrement dit $\overline{p \cup p'} = \emptyset$). Le cas n°3 correspond aux mêmes remarques pour H_J. Le cas n°4 se produit lorsque les remarques précédentes se produisent à la fois pour H_I et H_J.

D'autre part, par construction, l'une des deux classes p ou p' est non vide, (de même, q ou q'). Nous supposerons dans la suite qu'il s'agit de p et de q.

L'interaction $Int(p, p\prime, q, q\prime)$ entre $(p, p\prime)$ et $(q, q\prime)$ se définit, dans chacun de ces quatre cas, par :

a) **Cas n°1**: $Int(p, p\prime, q, q\prime) = (A_{p'q'} - A_{pq'}) - (A_{p'q} - A_{pq})$

b) **Cas n°2**: $Int(p, p\prime, q, q\prime) = A_{pq'} - A_{pq}$

c) **Cas n°3**: $Int(p, p\prime, q, q\prime) = A_{p'q} - A_{pq}$

d) **Cas n°4**: $Int(p, p\prime, q, q\prime) = A_{pq} - 1$

Nous détaillons ci-dessous les propriétés et l'interprétation de l'indice $Int(p, p\prime, q, q\prime)$ dans le cas 1 défini ci-dessus. Les cas suivants présentent des propriétés analogues. Elles seront résumées au paragraphe 3.3.

3.2. Interaction entre deux couples (p,p') et (q,q') vérifiant le cas n°1

Dans le cas n°1, cette interaction $Int(p, p\prime, q, q\prime)$ est définie par : $(A_{p'q'} - A_{pq'}) - (A_{p'q} - A_{pq})$.

Différentes propriétés ou remarques permettent d'interpréter cette définition.

PROPRIÉTÉ 5. — *La quantité* $\frac{1}{k} \cdot \sqrt{\frac{k(p).k(p')}{k(p)+k(p')} \cdot \frac{k(q).k(q')}{k(q)+k(q')}} . Int(p, p\prime, q, q\prime)$ *s'interpréte comme un coefficient de corrélation*

Démonstration. — On introduit deux variables U et V définies sur l'ensemble L des k individus (répartis dans les cases du tableau k_{IJ}) : $\forall \ell \in L$,

$$U(\ell) = \begin{cases} \frac{1}{k(p)} \text{ si } \ell \in p \\ 0 \text{ si } \ell \in \overline{p \cup p'} \\ -\frac{1}{k(p')} \text{ si } \ell \in p' \end{cases} \qquad V(\ell) = \begin{cases} \frac{1}{k(q)} \text{ si } \ell \in q \\ 0 \text{ si } \ell \in \overline{q \cup q'} \\ -\frac{1}{k(q')} \text{ si } \ell \in q' \end{cases}$$

$\overline{p \cup p'}$ (respectivement $\overline{q \cup q'}$) représente l'ensemble des éléments de L n'appartenant pas à $p \cup p'$ (resp. à $q \cup q'$).

De simples calculs montrent que U et V sont centrées et que leur corrélation est égale à la quantité donnée ci-dessus.

Remarques. —

a) Ce coefficient de corrélation $cor(U, V)$ prend respectivement les valeurs 1 et -1 lorsque le tableau de contingence 3×3 croisant $(p, p\prime, \overline{p \cup p'})$ et $(q, q\prime, \overline{q \cup q'})$ prend les formes suivantes :

	q	$q\prime$	$\overline{q \cup q'}$
p	a	0	0
$p\prime$	0	b	0
$\overline{p \cup p'}$	0	0	c
$Cor(U, V) = +1$			

	q	$q\prime$	$\overline{q \cup q'}$
p	0	a	0
$p\prime$	b	0	0
$\overline{p \cup p'}$	0	0	c
$Cor(U, V) = -1$			

b) $Cor(U, V) = 0$ si l'on a $A_{pq} + A_{p\prime q\prime} = A_{p\prime q} + A_{pq\prime}$

Ce cas inclut celui de l'indépendance.

Le modèle hypergéométrique 3×3

Le coefficient de corrélation $cor(U,V)$ ne dépend que des valeurs du tableau de contingence 3×3 croisant $(p, p\prime, \overline{p \cup p'})$ et $(q, q\prime, \overline{q \cup q'})$.

	q	$q\prime$	$\overline{q \cup q'}$	**marge**
p	w	u		$k(p)$
$p\prime$	t	v		$k(p\prime)$
$\overline{p \cup p'}$				$k - k(p) - k(p\prime)$
marge	$k(q)$	$k(q')$	$k - k(q) - k(q')$	

Les cases non remplies du tableau 3×3 croisant $(p, p\prime, \overline{p \cup p'})$ et $(q, q\prime, \overline{q \cup q'})$ ont leurs contenus calculés à partir des marges et des quatre valeurs u,v,w,t.

Le modèle hypergeométrique suppose que les éléments de L sont répartis au hasard dans les cases de ce tableau 3×3, les marges étant fixées. L'hypothèse H_0 d'un modèle hypergéométrique traduit l'absence de liens entre les lignes et colonnes de ce tableau 3×3.

Cette loi hypergéométrique est définie par les probabilités :

$P(u, v, w, t) = P(k(p, q') = u, k(p', q') = v, k(p, q) = w, k(p', q) = t)$

La définition de cette loi et de ses moments sont bien connues (voir par exemple Plackett, 1981 ou Lancaster, 1969). On déduit (voir Denimal-Camiz 2001) :

Propriété 6. — *Sous cette hypothèse H_0 d'un modèle hypergéométrique, l'espérance et la variance de la statistique $Int(p, p\prime, q, q\prime)$ valent :*

$E\left[Int(p, p\prime, q, q\prime)\right] = 0$

$$Var\left[Int(p, p\prime, q, q\prime)\right] = \frac{k^2}{k-1} . \frac{k(p) + k(p')}{k(p).k(p')} . \frac{k(q) + k(q')}{k(q).k(q')}$$

Conséquence. — *On vérifie que :* $\dfrac{Int(p, p\prime, q, q\prime)}{\sqrt{Var\left[Int(p, p\prime, q, q\prime)\right]}} = \sqrt{k-1}.cor(U, V)$

Test conditionnel exact d'interaction

Le test conditionnel exact bati pour détecter une liaison significative entre deux couples $(p, p\prime)$ et $(q, q\prime)$ est basé sur la statistique $Int(p, p\prime, q, q\prime)$. L'hypothèse H_0 nulle testée est l'hypothèse du modèle hypergéométrique précédent. La valeur observée de $Int(p, p\prime, q, q\prime)$ étant noté Int_{obs}, et en supposant par exemple $Int_{obs} \geqslant 0$, on rejettera l'hypothèse H_0 traduisant l'absence de liens entre $(p, p\prime)$ et $(q, q\prime)$ si la p-valeur $P\left[Int(p, p\prime, q, q\prime) \geqslant int_{obs}\right]$ est plus petite qu'un risque de première espèce donné.

3.3. Interaction entre deux couples (p,p') et (q,q') vérifiant les cas autres que 1

Le test conditionnel exact d'interaction entre (p, p') et (q, q') dans les autres cas est également basé sur la statistique $Int(p, p\prime, q, q\prime)$ et l'hypothèse d'absence de liens entre les deux couples (p, p') et (q, q') est encore défini par un modèle hypergéométrique. Seuls, les dimensions du tableau de contingence associé diffèrent.

Nous résumons dans le tableau suivant les caractéristiques du test :

Cas	**Modèle hypergéométrique (H_0)**	**Statistique T du test**
2	*modèle (2,3)*	$A_{pq'} - A_{pq}$
3	*modèle (3,2)*	$A_{p'q} - A_{pq}$
4	*modèle (2,2)*	$A_{pq} - 1$

Cas	$\mathbf{E_{H_0}(T)}$	$\mathbf{Var_{H_0}(T)}$
2	*0*	$\frac{k}{k-1}.\frac{k-k(p)}{k(p)}.\frac{k(q)+k(q')}{k(q).k(q')}$
3	*0*	$\frac{k}{k-1}.\frac{k-k(q)}{k(q)}.\frac{k(p)+k(p')}{k(p).k(p')}$
4	*0*	$\frac{1}{k-1}.\frac{k-k(p)}{k(p)}.\frac{k-k(q)}{k(q)}$

Dans le cas du modèle (2,2), on retrouve la loi hypergéométrique classique.

3.4. Calcul des p-valeurs des tests conditionnels exacts d'interaction

Le calcul de cette p-valeur repose sur la loi hypergéométrique multiple. Ce calcul est réalisé par un tirage de n tableaux de contingence de marges fixées par un algorithme dû à Patefield (1981). Une approximation de la p-valeur est ainsi obtenue à partir de l'échantillon tiré. L'algorithme de Patefield est rapide et présente également l'avantage de fournir la probabilité du tableau généré.

Dans le cas de l'exemple traité dans cet article, la taille n de l'échantillon a été choisi à 10000.

3.5. Test entre deux nœuds $n \in H_I$ et $m \in H_J$

Pour chacune des deux hiérarchies H_I et H_J,chaque nœud obtenu est en fait la fusion de deux dipoles.

Ainsi, un nœud n de H_I est la fusion des deux dipoles (p_1, p_2) et (p'_1, p'_2) telles que $p_1 \subset I$, $p_2 \subset I$, $p'_1 \subset I$, $p'_2 \subset I$ et les quatre classes p_1, p'_1, p_2, p'_2 sont disjointes 2 à 2.

Le dipole associé au nœud n s'interpréte comme le compromis des deux dipoles (p_1, p_2) et (p'_1, p'_2).

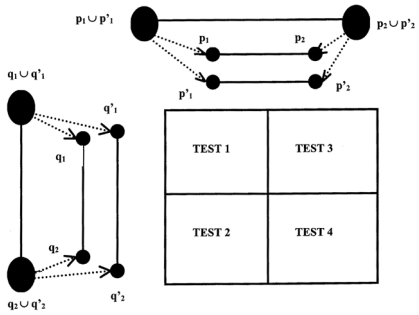
Généralement, le dipôle associé au nœud n est $(p_1 \cup p'_1, p_2 \cup p'_2)$

De même, un nœud m de H_J est la fusion de deux dipoles (q_1, q_2) et (q'_1, q'_2).

L'interaction entre les deux nœuds n et m sera d'autant plus élevée si l'écart entre les nouvelles classes formées au nœud n s'explique par celui observé entre les nouvelles classes formé au nœud m.

En conséquence, une interaction significative entre n et m est observée si l'une au moins des quatre interactions possibles entre les quatre paires de couples suivants est également significative :

Tests	*couple 1*	*couple 2*
1	(p_1, p'_1)	(q_1, q'_1)
2	(p_1, p'_1)	(q_2, q'_2)
3	(p_2, p'_2)	(q_1, q'_1)
4	(p_2, p'_2)	(q_2, q'_2)



L'interaction entre deux couples est réalisée à partir du test conditionnel exact d'interaction exposé précédemment.

Il est possible par exemple que l'une des classes p_1 ou p'_1 soient vides. Dans ce cas, seules deux interactions seront à tester.

3.6. Élagage de H_I et H_J

Un nœud n de H_I est dit significatif s'il existe au moins un nœud m de H_J ayant une interaction significative avec n. Une définition analogue permet d'identifier les nœuds de H_J significatifs.

Deux règles de coupure de chacune des hiérarchies H_I et H_J sont possibles : soit la coupure est définie la plus haute possible de façon à ce que les nœuds qui lui soient inférieurs soient non significatifs, soit la coupure est définie la plus basse possible de façon à ce que tous les nœuds qui lui soient supérieurs soient significatifs. Dans l'exemple traité dans cet article, la seconde option a été choisie.

4. Aides à l'interprétation des hiérarchies sur I et sur J

On considère les hiérarchies optimisées et élaguées H_I et H_J. Chacune d'elles est composée d'un ensemble de nœuds ou dipoles supérieurs et d'un ensemble de dipoles terminaux. Les classes terminales composant ces derniers forment une partition de l'ensemble I ou J.

Les tests précédents permettent de d'identifier et d'expliquer les interactions significatives entre les nœuds de H_I et H_J.

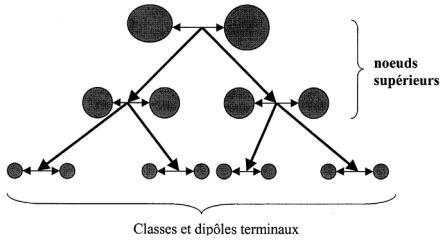


FIG 2. — Un exemple de hiérarchie optimisée.

4.1. Étiquetage d'une hiérarchie à partir des dipoles terminaux de l'autre

Des tests complémentaires sont ici proposés de façon à identifier les classes terminales de la partition obtenue sur I (respectivement sur J) ayant une interaction significative avec un nœud donné de l'autre hiérarchie H_J (respectivement H_I).

Soit $n = (p_1, p_2)$ un nœud de H_I .et soit un dipole terminal (q_1, q_2) de H_J

Un test conditionnel exact d'interaction entre (q_1, q_2) et (p_1, p_2) sera mené afin de vérifier la significativité de cette interaction.

Ainsi, il est possible d'expliquer les nœuds et dipoles terminaux de H_I à partir des dipoles terminaux de H_J, et par suite d'étiqueter la hiérarchie H_I à partir des dipoles terminaux de H_J.

La même interprétation peut être aussi faite en permutant les rôles des hiérarchies H_I et H_J.

4.2. Représentations factorielles associées à chaque nœud

Considérons par exemple le nœud $n = (p_1, p_2)$ de H_I. Par construction, il est issu d'une ACP d'un tableau à 2 colonnes ou d'une analyse des correspondances (AFC) équivalente d'un tableau de type $K\,[c_1, \overline{c_1}, c_2, \overline{c_2}]$.

Considérons cette dernière analyse et adjoignons à ce tableau en colonnes supplémentaires les lignes du tableau initial k_{IJ} correspondant aux contenus des classes p_1 et p_2 :

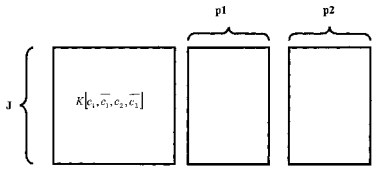


FIG 3. — Analyse des correspondances associée au nœud n.

Des formules de transition distribuent les éléments de J suivant les quatre points $c_1, \overline{c_1}, c_2, \overline{c_2}$ *et de même les éléments des classes* p_1, p_2 *se répartissent suivant les points représentant les éléments de J par d'autres formules de transition. Dans l'exemple traité ci-dessous, deux représentations factorielles planes seront données pour chaque nœud étudié : l'une représente les éléments de J et les quatre points* $c_1, \overline{c_1}, c_2, \overline{c_2}$ *et l'autre les élements des classes* p_1 *et* p_2.

5. Application à un exemple

La méthodologie est appliquée à un tableau de contingence croisant l'ensemble I des 96 *départements français et l'ensemble J des candidats à l'élection présidentielle de* 1995. *Le tableau peut être trouvé dans le livre «l'analyse des données évolutives» (Dazy, Le Barzic, 1996).*

Les résultats obtenus peuvent se ranger en 4 grandes étapes : construction des hiérarchies initiales sur I et sur J et leur optimisation (étapes 1 et 2), élagages des hiérarchies (étape 3) et interprétation des hiérarchies (étape 4).

D'aprés la propriété 4, en considérant par exemple la hiérarchie initiale H_J *construite sur J, la somme des indices de niveau des nœuds de la hiérarchie initiale augmentée de la variance de la variable représentative totale est égale à la somme des variances des initiales* y^j, $j \in J$. *Cette dernière quantité est appelée, en bref, dans la suite «inertie totale». Le logiciel mettant en œuvre cette méthodologie, offre la possibilité de réaliser l'étape d'optimisation et par suite celle de l'élagage en ne considérant pas les nœuds les plus bas de la hiérarchie initiale et en se limitant aux r nœuds les plus hauts conduisant, par exemple, à* 90% *de l'inertie totale :*

$$\sum_{k=|J|-1-r}^{|J|-1} \nu(n_k) + var(\sum_{j \in J} a_j y^j) = 90\% \text{ de l'inertie totale}$$

Cette modification permet de limiter le temps de calcul pour les étapes d'optimisation et d'élagage tout en limitant la possibilité d'exclure un nœud significatif.

Cette approche a été appliquée à l'exemple traité. Dans le cadre de l'exemple traité, pour chacune des deux hiérarchies initiales, l'amélioration du critère à maximiser, dans la phase d'optimisation, reste inférieur au seuil fixé par le programme à savoir 0.3%. En conséquence, les hiérarchies initiales limitées aux nœuds les plus hauts, à savoir à ceux conduisant à 90% de l'inertie totale, seront considérées comme optimales et soumises directement à l'étape d'élagage.

Nous présentons ci-dessous aux paragraphes 5.1 *et* 5.2 *les tableaux décrivant les deux hiérarchies initiales* H_I *et* H_J.

5.1. Étape 1 : Hiérarchie initiale H_J

Le tableau 1 rassemble les coefficients a_j définissant la variable représentative totale $\sum_{j\in J} a_j.y^j$. On peut ainsi remarquer l'importance du coefficient associé au candidat Lepen. On vérifiera, plus loin, l'importance de ce dernier dans l'interprétation de H_J.

TABLEAU 1. — Coefficients $\left(a_j\right)_{j\in J}$.

Modalités	Abst.	Blancs	DeVilliers	Lepen	Chirac	Laguillier
Numérotation	1	2	3	4	5	6
a_j	0.07	-0.03	-0.12	0.79	-0.56	-0.01

Cheminade	Jospin	Voynet	Balladur	Hue
7	8	9	10	11
0.00	-0.07	-0.03	-0.10	0.14

Le tableau 2 décrit la hiérarchie initiale H_J. Les nœuds de H_J sont numérotés de $card(J)+1$ à $2.card(J)-1$. Les trois premières colonnes indiquent les numéros des nœuds n et leurs ainés et benjamins $a(n)$ et $b(n)$. Chaque nœud est issu de l'ACP d'un tableau à 2 colonnes y^q et $y^{q'}$ ou de l'analyse des correspondances équivalente (propriété 3) d'un tableau noté $K[q,\overline{q},q',\overline{q'}]$. Les deux valeurs propres obtenues sont notées $\lambda_1 \geqslant \lambda_2$ et représentent les colonnes 6 et 7. Les deux colonnes 4 et 5 donnent les pourcentages $\%\lambda_1 = \dfrac{\lambda_1 * 100}{\lambda_1+\lambda_2}$ et $\%\lambda_2 = \dfrac{\lambda_2 * 100}{\lambda_1+\lambda_2}$. Enfin, suivant la propriété 4, les deux dernières colonnes donnent le pourcentage de chaque seconde valeur propre par rapport à la quantité appelée ci-dessus « inertie totale », ainsi que le pourcentage cumulé associé.

TABLEAU 2. — Résultats de la hiérarchie initiale H_J.

noeud	ainé	benjamin	%lamda1	%lamda2	lamda1	lamda2	%inertie	%cumulée
12	2	7	98.4353	1.5647	0.0012	0.0000	0.0345	0.0345
13	1	12	92.0716	7.9284	0.0054	0.0005	0.8782	0.9128
14	6	8	89.0793	10.9207	0.0050	0.0006	1.1520	2.0647
15	9	11	89.8936	10.1064	0.0076	0.0009	1.6009	3.6656
16	10	15	82.6748	17.3252	0.0103	0.0022	4.0824	7.7480
17	13	14	73.1341	26.8659	0.0076	0.0028	5.2644	13.0125
18	4	5	81.4364	18.5636	0.0161	0.0037	6.9255	19.9379
19	17	3	62.3697	37.6303	0.0100	0.0060	11.3643	31.3022
20	19	16	57.5413	42.4587	0.0117	0.0086	16.2687	47.5709
21	20	18	58.9495	41.0505	0.0164	0.0114	21.5224	69.0933
total						0.0367	69.0933	

Le tableau 3 présente la somme des secondes valeurs propres (total lambda2), la variance de la variable représentative totale $\sum_{j\in J} a_j y^j$ (Voir §2.5, avant la propriété 4) ainsi que leur somme (propriété 4) appelée ci-dessus « inertie totale » et représentant la somme des variances de l'ensemble des variables y^j, $j \in J$.

TABLEAU 3. — Décomposition de l'inertie totale (propriété 4).

Total lambda2	0.0367
Variance de la variable représentative de la classe complète	0.0164
Somme des variances des 11 variables initiales	0.0531

5.2. Étape 2 : Hiérarchie initiale H_I

Des tableaux identiques notés 4,5 et 6 caractérisent la hiérarchie initiale H_I. Seuls les coefficients a_i supérieurs en valeur absolue à 0.10 sont cités au tableau 4. Les départements Vendée, Bas Rhin, Correze, Haut Rhin auront des votes caractéristiques comme nous le verrons plus loin.

TABLEAU 4. — Coefficients a_i les plus importants.

Modalités i	Vendée(85)	BasRhin(67)	Correze(19)	HautRhin(68)	Alpes maritimes(6)
a_i	-0.36	0.32	-0.28	0.24	0.21

Moselle(57).	HauteVienne(87)	Dordogne(24)	Var(83)	Haute Garonne(31)	Seine StLouis(93)
0.21	-0.20	-0.17	0.17	-0.15	0.15

BouchesduRhone(13)	Cotesd'armor(22)	Rhone(69)	Cantal(15)	MaineetLoire(49)	Landes(40)
0.14	-0.14	0.14	-0.13	-0.13	-0.12

Paris(75)	Vaucluse(84)
-0.11	0.11

Le tableau 5 décrit les nœuds les plus hauts de la hiérarchie initiale H_I conduisant à 90% de «l'inertie totale».

TABLEAU 5. — Résultats de la hiérarchie initiale H_I.

noeud	ainé	benjamin	%lamda1	%lamda2	lamda1	lamda2	%inertie	%cumulée
168	127	156	93.3703	6.6297	0.0023	0.0002	0.3475	7.1481
169	152	158	94.1891	5.8109	0.0029	0.0002	0.3798	7.5279
170	165	157	92.4070	7.5930	0.0022	0.0002	0.3822	7.9101
171	137	161	95.8579	4.1421	0.0043	0.0002	0.3886	8.2987
172	141	162	93.2220	6.7780	0.0030	0.0002	0.4515	8.7502
173	159	154	93.7472	6.2528	0.0033	0.0002	0.4674	9.2175
174	149	167	89.0867	10.9133	0.0019	0.0002	0.4776	9.6951
175	38	153	81.2659	18.7341	0.0010	0.0002	0.4863	10.1814
176	164	169	93.2616	6.7384	0.0033	0.0002	0.4932	1.6746
177	163	147	95.1432	4.8568	0.0051	0.0003	0.5455	11.2202
178	160	134	85.2394	14.7606	0.0017	0.0003	0.6139	11.8341
179	166	144	94.7019	5.2981	0.0058	0.0003	0.6814	12.5155
180	173	175	92.1993	7.8007	0.0040	0.0003	0.7111	13.2267
181	170	179	94.4377	5.5623	0.0076	0.0004	0.9362	14.1629
182	176	178	90.5704	9.4296	0.0045	0.0005	0.9780	15.1409
183	155	172	88.5454	11.4546	0.0040	0.0005	1.0939	16.2348
184	168	174	86.5051	13.4949	0.0036	0.0006	1.1867	17.4215
185	180	183	84.5126	15.4874	0.0068	0.0012	2.6114	20.0328
186	181	182	89.4404	10.5596	0.0108	0.0013	2.6704	22.7033
187	184	177	81.1371	18.8629	0.0071	0.0016	3.4507	26.1540
188	171	186	82.4660	17.5339	0.0124	0.0026	5.5429	31.6969
189	86	185	78.7193	21.2807	0.0103	0.0028	5.8290	37.5258
190	187	189	61.5235	38.4765	0.0107	0.0067	14.0073	51.5332
191	188	190	59.5958	40.4042	0.0138	0.0093	19.5826	71.1158
total						0.0339	71.1158	

TABLEAU 6. — Décomposition de l'inertie totale (propriété 5).

Total lamda2	0.0339
Variance de la variable représentative de la classe complète	0.0138
Somme des variances des 96 variables initiales	0.0477

5.3. Étape 3 : Élagages mutuels des hiérarchies optimisées

Dans le cas de notre exemple, les hiérarchies H_I et H_J sont d'abord réduites aux nœuds les plus hauts conduisant à 90 % de l'inertie totale. Puis, la procédure d'élagage est appliquée. La hiérarchie H_J (respectivement H_I) est ainsi réduite avant élagage aux 5 nœuds (respectivement 30 nœuds) les plus hauts.

Selon le paragraphe 3.5, pour deux nœuds n et m appartenant respectivement aux deux hiérarchies H_I et H_J, quatre tests conditionnels exacts au maximum peuvent être menés. L'association entre ces deux nœuds n et m est dite significative si l'un au moins de ces quatre tests révèle une interaction significative. Le tableau 7 rassemble, pour chaque couple de nœuds retenus, le minimum des quatre p-valeurs associées à ces quatre tests.

Un nœud n de H_I est déclaré significatif (noté S dans le tableau 7) s'il existe au moins un nœud m de H_J pour lequel l'association avec n est significative. Dans le cas contraire, le nœud n est déclaré non significatif (noté NS dans le tableau 7). Les nœuds significatifs de H_J sont définis de manière analogue.

Dans le tableau 7, les nœuds significatifs de H_I et H_J ont été identifiés à partir du seuil $\alpha = 0.001$.

La règle choisie pour élaguer H_I ou H_J est de couper chaque branche de l'arbre le plus bas possible de façon à ce que les nœuds supérieurs à cette coupure soient tous significatifs. En conséquence, les 5 nœuds retenus de H_J sont conservés et les 30 nœuds retenus de H_I sont élagués comme indiqués au tableau 8 ci-dessous.

TABLEAU 7. — p-valeurs.

Noeuds H_J				21	20	19	18	17
Aîné				20	19	17	4	13
Benjamin				18	16	3	5	14
Noeuds H_I	Aîné	Benj.	sign.	S	S	S	S	S
191	188	190	S	0.000	0.0000	0.0001	0.0000	0.0000
190	187	189	S	0.0000	0.0000	0.0006	0.0001	0.0215
189	86	185	S	0.0000	0.0001	0.0000	0.0011	0.0211
188	171	186	S	0.0000	0.0003	0.0001	0.0000	0.0000
187	184	177	NS	0.0660	0.0025	0.0139	0.0610	0.0354
186	181	182	S	0.0000	0.0001	0.0000	0.0000	0.0000
185	180	183	S	0.0000	0.0010	0.0188	0.0003	0.0203
184	168	174	NS	0.0977	0.0041	0.1123	0.1015	0.3663
183	155	172	S	0.0006	0.0199	0.0045	0.0007	0.0554
182	176	178	S	0.0074	0.0010	0.0005	0.0026	0.0013
181	170	179	S	0.0003	0.0008	0.0746	0.0008	0.0557
180	173	175	S	0.0001	0.0011	0.0258	0.1008	0.0930
179	166	144	S	0.0004	0.0014	0.0954	0.0001	0.0347
178	160	134	NS	0.1115	0.0553	0.0293	0.2341	0.0036
177	163	147	NS	0.0290	0.0119	0.0153	0.0358	0.0452
176	164	169	S	0.0082	0.0032	0.0021	0.0008	0.0019
175	38	153	NS	0.1203	0.3188	0.0240	0.2738	0.0552
174	149	167	NS	0.1122	0.0050	0.1022	0.2265	0.1257
173	159	154	NS	0.0053	0.0171	0.0066	0.1646	0.0511
172	141	162	NS	0.0044	0.0604	0.0807	0.0087	0.0498
171	137	161	S	0.0001	0.0751	0.1264	0.0001	0.1572
170	165	157	NS	0.0120	0.2012	0.1532	0.0037	0.1888
169	152	158	NS	0.0381	0.0243	0.0233	0.0411	0.0229
168	127	156	NS	0.0125	0.0058	0.1943	0.0346	0.3545
167	117	151	NS	0.2426	0.0924	0.0783	0.1502	0.1728
166	136	150	S	0.0004	0.0014	0.0600	0.0001	0.2520
165	58	148	NS	0.0176	0.2567	0.0875	0.0035	0.0560
164	115	135	NS	0.1642	0.1165	0.0681	0.2463	0.0942
163	76	110	NS	0.0001	0.0131	0.0344	0.0003	0.0641
162	13	145	NS	0.0195	0.0730	0.0980	0.0133	0.2763

TABLEAU 8. — Élagage de H_I.

Noeuds conservés de la hiérarchie H_I	Noeuds élagués de la hiérarchie H_I
191 190 187	184 168 127
191 190 187	184 168 156
191 190 187	184 174 149
191 190 187	184 174 167 117
191 190 187	184 174 167 151
191 190 187	177 163 76
191 190 187	177 163 110
191 190 189	177 147
191 190 189 86	
191 190 189 185 180 173	159
191 190 189 185 180 173	154
191 190 189 185 180 175	38
191 190 189 185 180 175	153
191 190 189 185 183 155	
191 190 189 185 183 172	141
191 190 189 185 183 172	162 13
191 188 186 185 183 172	162 145
191 188 186 181 170	165 58
191 188 186 181 170	165 148
191 188 186 181 170	157
191 188 186 181 179 166 136	
191 188 186 181 179 166 150	
191 188 186 179 144	
191 188 186 182 176 164	115
191 188 186 182 176 164	135
191 188 186 182 176 169	152
191 188 186 182 176 169	158
191 188 186 182 178	160
191 188 171 182 178	134
191 188 171 137	
191 188 181 161	

On trouvera ci-dessous la représentation des hiérarchies élaguées H_I et H_J (Figures 4 et 5) et le contenu de leurs classes terminales (Tableaux 9 et 10). Comme précisé au début du § 3, chaque nœud est un dipôle composé de deux

ensembles de modalités. Dans les tableaux 9 et 10, pour chaque nœud n*, ces deux ensembles seront noté* n_1 *et* n_2 *.*

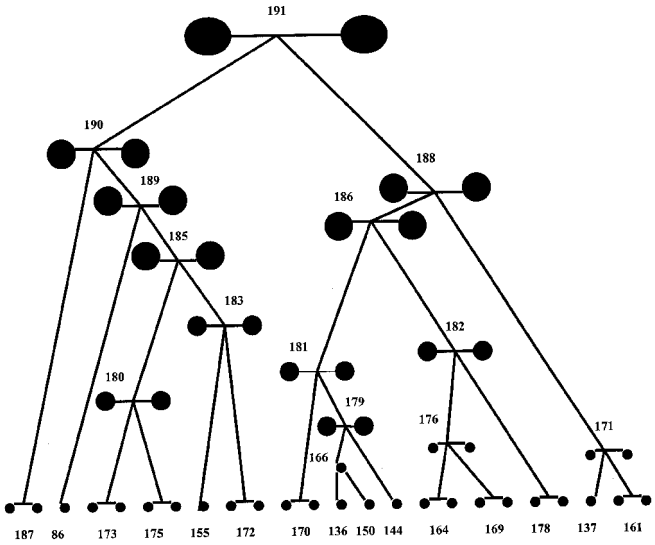


FIG 4. — Représentation de H_I.

TABLEAU 9. — Contenus des classes terminales de la hiérarchie H_I.

	Contenus des classes terminales de H_I		Contenus des classes terminales de H_I
187-1	5 20A 20B 21 25 50 75 78 92	**136**	55 67 68 69
187-2	3 4 18 30 34 39 47 59 62 66 76 80	**150**	1 10 28 45 52 88 89
86	85	**144**	24 46 87
173-1	43 49 53 56 61	**164-1**	7 70 81 82
173-2	93 95	**164-2**	77
175-1	37 41	**169-1**	22 40 63 65
175-2	91 94	**169-2**	6 51 83
155	17 44 72 79	**178-1**	9 11 31 33 58
172-1	12 14 29 35 48 64	**178-2**	74
172-2	8 13	**137**	38 54 90
170-1	16 32 36 71 86	**161-1**	15 19 23
170-2	27 42 57 60 84	**161-2**	26 73

TABLEAU 10. — Contenus des classes terminales de H_J.

	Contenus des classes terminales de H_J
13-1	Blancs (2) - Cheminade (7)
13-2	Abstentions (1)
14	Laguillier (6) - Jospin (8)
3	De Villiers (3)
16-1	Voynet (9) - Balladur (10)
16-2	Hue (11)
4	Lepen (4)
5	Chirac (5)

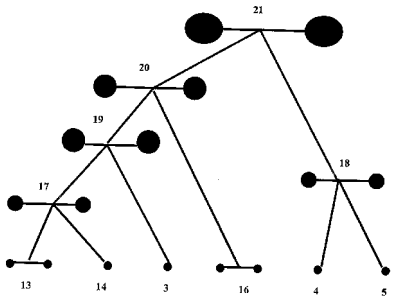


FIG 5. — Représentation de la hiérarchie H_J.

5.4. Étape 4 : Interprétations mutuelles des hiérarchies optimisées et élaguées

On choisit ici d'interpréter nœuds et classes terminales de H_I en fonction des classes terminales de H_J. Chaque nœud de H_I sera étiqueté suivant les dipoles terminaux de H_J jouant un rôle significatif dans l'interpétation de ce nœud. Le choix contraire, obtenu en permutant les rôles des hiérarchies H_I et H_J, aurait pu être fait. Il aurait généré la même information répartie dans ce cas sur les nœuds de H_J.

Le tableau 11 rassemble les p-valeurs associées aux tests d'interaction entre les dipoles terminaux de H_J et les dipoles supérieurs de H_I. Le seuil de $\alpha = 1\%$ ayant été choisi, les p-valeurs du tableau 11 traduisant une interaction significative ont été inscrites en caractères gras. Le tableau 12 donne pour ces interactions significatives les quantités $A_{pq} = \dfrac{k.k(p,q)}{k(p).k(q)}$ (§3.1) entre les classes p et q concernées.

Les tableaux 13 et 14 ont le même signification mais pour les interactions entre les dipoles terminaux des deux hiérarchies H_I et H_J.

L'ensemble des résultats contenus dans ces tableaux 11 à 14 permet d'étiqueter la hiérarchie H_I. (Figures 6,7 et 8). De plus, à chaque nœud de H_I, deux représentations factorielles selon le §4.2 permettent de visualiser les correspondances entre les départements constituant le dipôle associé au nœud et l'ensemble des candidats (Figures 9 à 14).

On résume ci-dessous l'ensemble de ces résultats.

Le nœud 191, sommet de la hiérarchie, représente un dipôle constitué de deux classes de départements dont les différences de votes s'expliquent par l'opposition entre le candidat LePen d'une part et les candidats Jospin, Chirac, Devilliers d'autre part. On peut remarquer que le candidat LePen joue un rôle important dans l'explication des différences de votes entre départements. En examinant les représentations factorielles 9 et 10, on peut remarquer les départements Bas Rhin et Haut Rhin (67 et 68) sont ceux qui totalisent des votes élevées pour le candidat LePen. La scission du nœud 191 en les deux nœuds 190 et 188 fait ensuite intervenir le candidat Hue.

Dans le cas du nœud 190, les deux groupes de départements associés s'opposent l'un par des votes plus élevés pour les candidats Hue et Lepen (autrement dit pour les extrêmes) l'autre par des votes plus élevés pour la droite classique (Chirac, Balladur, Devilliers). En examinant les représentations factorielles 13 et 14, on vérifie que le vote plus élevé pour les extrêmes s'observe pour les Bouches du Rhone (13) ou pour La Seine St Denis (93), alors que les départements Deux Sèvres (79) ou Maine et Loire(49) ont les caractéristiques opposées. La Vendée(85) trés éloignée s'explique par ses votes trés élevés pour De Villiers.

Dans le cas du nœud 188, le candidat Lepen s'oppose au candidat Hue ainsi qu'aux candidats Chirac, Jospin Laguillier. En examinant les représentations factorielles 11 et 12, on retrouve les départements Haut Rhin et Bas Rhin (67 et 68) associés à Lepen. On observe, de même, la Corrèze (19) associée à Chirac et Hue ainsi que le Gers (32) admettant des votes à gauche plus élevés (Jospin, Hue, Laguiller). Le nœud 188 se rescinde ensuite en 171 et 186, le premier s'expliquant par une opposition plus marquée entre Lepen et Chirac et le second entre Lepen et la gauche (Jospin, Hue, Laguiller).

Cette description des premiers nœuds de la hiérarchie peut encore être affinée par l'étude des nœuds suivants en procédant de manière analogue.

6. Conclusion

Il convient d'insister sur la qualité des hiérarchies obtenues garantie par les étapes d'optimisation et d'élagage et sur la visualisation possible des correspondances entre éléments de I et de J par les différents plans factoriels obtenus. D'autre part, la méthode se distingue également par le fait que les nœuds des hiérarchies soient des dipôles de modalités. Il apparaît, dans l'exemple traité, que cette approche en dipôles permet de mettre en lumière une information qui serait moins facilement révélée par les classifications classiques. Enfin, cet article a été rédigé en prévision de sa généralisation au cas des correspondances multiples.

TABLEAU 11. — p-valeurs des tests d'interaction entre les dipôles supérieurs de H_I et les dipôles terminaux de H_J.

	Dipôles terminaux de la hiérarchie H_J					
Dipoles supérieurs de H_I	13	14	3	16	4	5
191	1.41%	**0.02%**	**0.90%**	42.6%	**0.00%**	**0.57%**
190	3.65%	24.2%	**0.01%**	**0.00%**	**0.14%**	18.4%
189	2.10%	8.30%	**0.05%**	**0.00%**	**0.08%**	22.8%
188	8.05%	**0.00%**	40.0%	**0.05%**	**0.00%**	**0.98%**
186	9.19%	**0.00%**	41.3%	**0.12%**	**0.00%**	13.1%
185	2.1%	7.1%	**0.87%**	**0.00%**	**0.12%**	21.6%
183	21.8%	5.8%	15.2%	1.4%	**0.02%**	10.1%
182	12.5%	**0.00%**	40.1%	7.9%	**0.14%**	39.4%
181	10.2%	2.59%	33.1%	**0.45%**	**0.00%**	2.45%
180	1.9%	39.8%	1.7%	**0.08%**	21.2%	48.9%
179	6.8%	5.2%	34.2%	**0.07%**	**0.06%**	1.88%
176	9.0%	**0.15%**	44.8%	24.5%	**0.28%**	49.2%
171	57.4%	36.1%	28.9%	5.4%	**0.78%**	**0.00%**
166	36.1%	3.2%	24.3%	**0.14%**	**0.00%**	11.6%

TABLEAU 12. — Quantités A_{pq} entre les classes p et q de H_I (dipôles supérieurs) et H_J (dipôles terminaux).

	Dipôles terminaux de la hiérarchie H_J					
Dip. sup. de H_I	**14**	**3**	**16-1**	**16-2**	**4**	**5**
(191-1;191-2)	(1.08;0.93)	(1.14;0.88)			(0.76;1.19)	(1.07;0.94)
(190-1;1910-2)		(1.36;0.78)	(1.15;0.88)	(0.80;1.24)	(0.77;1.02)	(1.05;0.98)
(189-1;189-2)		(1.47;0.73)	(1.18;0.82)	(0.81;1.26)	(0.72;1.10)	
(188-1;188-2)	(1.17;0.90)		(0.91;1.06)	(1.18;0.80)	(0.73;1.31)	(1.11;0.92)
(186-1;186-2)	(1.18;0.97)		(0.93;1.07)	(1.16;0.78)	(0.76;1.34)	
(185-1;185-2)		(1.27;0.73)	(1.18;0.82)	(0.82;1.26)	(0.73;1.10)	
(183-1;183-2)					(0.67;1.35)	
(182-1;182-2)	(1.22;0.79)				(0.82;1.27)	
(181-1;181-2)			(0.90;1.09)	(1.27;0.76)	(0.82;1.27)	
(180-1;180-2)			(1.24;0.77)	(0.82;1.27)		
(179-1;179-2)			(0.76;1.14)	(1.68;0.69)	(0.60;1.35)	
(176-1;176-2)	(1.17;0.79)				(.0.78;1.30)	
(171-1;171-2)					(0.40;1.18)	(2.20;0.85)
166			1.14	0.70	1.35	

TABLEAU 13. — p-valeurs des tests d'interaction entre les dipôles terminaux de H_I et H_J.

	Dipôles terminaux de la hiérarchie H_J					
Dipoles terminaux de H_I	13	14	3	16	4	5
187	5.6%	15.9%	25.0%	**0.00%**	**0.12%**	**0.01%**
86	3.4%	25.9%	**0.00%**	21.4%	5.7%	35.7%
173	9.0%	43.5%	9.4%	**0.12%**	9.7%	26.4%
175	15.3%	39.1%	2.4%	15.9%	36.3%	27.9%
155	10.3%	12.9%	**0.48%**	14.7%	**0.80%**	36.3%
172	27.9%	5.4%	41.2%	1.85%	**0.07%**	5.60%
170	13.1%	13.3%	24.2%	34.2%	**0.11%**	16.9%
136	50.5%	5.3%	41.5%	**0.14%**	**0.00%**	9.6%
150	24.8%	21.9%	5.6%	23.6%	2.7%	44.3%
144	3.6%	9.7%	15.1%	**0.75%**	2.12%	1.90%
164	19.4%	15.4%	41.9%	39.3%	28.8%	38.9%
169	6.83%	**0.18%**	36.5%	14.4%	**0.09%**	36.4%
178	51.2%	2.7%	16.2%	9.2%	20.2%	40.7%
137	53.9%	15.8%	27.9%	36.1%	11.4%	7.50%
161	35.8%	42.8%	12.5%	7.53%	1.5%	**0.00%**

TABLEAU 14. — Quantités Apq entre les classes p et q de H_I (dipôles terminaux) et H_J (dipôles terminaux).

	Dipôles terminaux de la hiérarchie H_J					
Dip. term. de H_I	**14**	**3**	**16-1**	**16-2**	**4**	**5**
(187-1;187-2)			(0.99;0.92)	(0.68;1.40)	(0.76;1.12)	(1.26;0.87)
86		5.2				
(173-1;173-2)			(1.29;0.72)	(0.76;1.35)		
(175-1;175-2)		(1.77;0.77)				
155		1.62				
(172-1;172-2)					(0.68;1.36)	
(170-1;170-2)					(0.80;1.43)	
136			1.22	0.61	1.48	
150					1.24	
144			0.76	1.69		
(164-1;164-2)						
(169-1;169-2)	(1.22;0.76)				(0.67;1.37)	
(178-1;178-2)						
137						
(161-1;161-2)						(2.20;0.92)

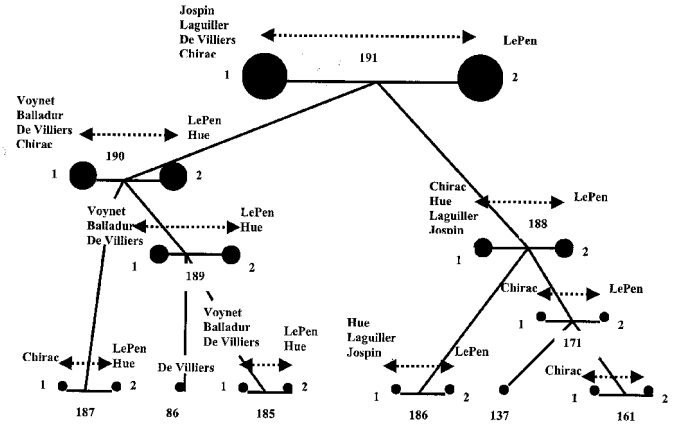


FIG 6. — Étiquetage de la hiérarchie H_I.

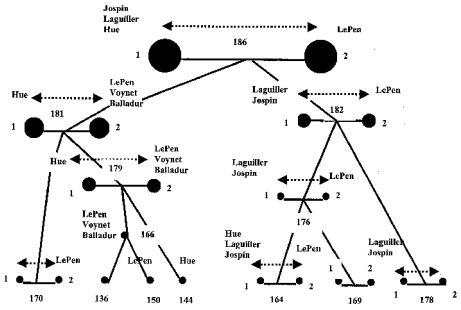


FIG 7. — Branche du nœud 186.

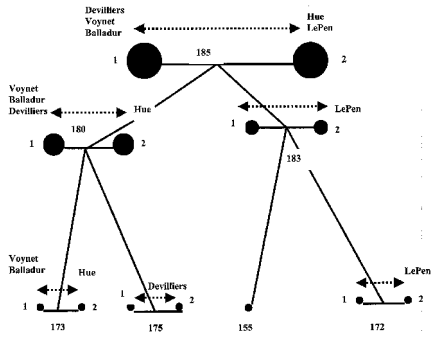


FIG 8. — Branche du nœud 185.

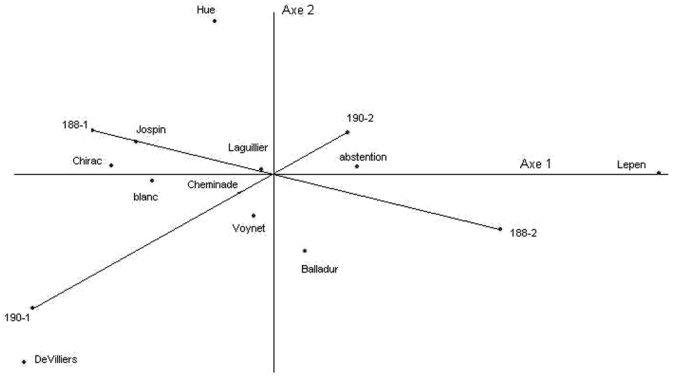


FIG 9. — Plan factoriel associé au nœud 191 d'ainé 188 et de benjamin 190.

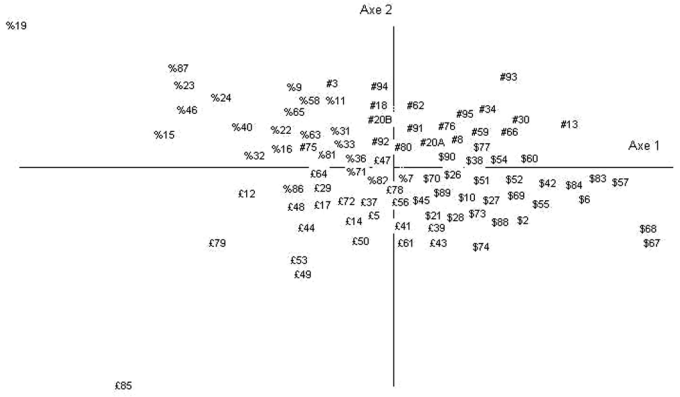


FIG 10. — Représentation des classes 188-1(%), 188-2($),190-1(£),190-2(#) issues du nœud 191.

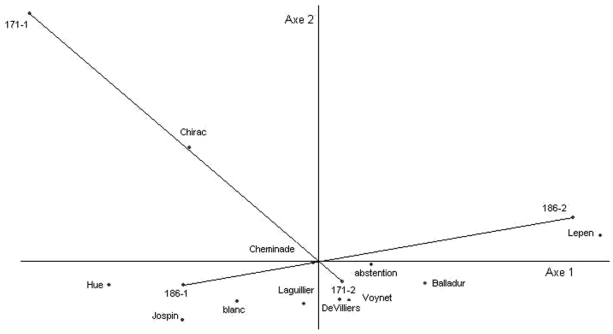


FIG 11. — Plan factoriel associé au nœud 188 d'ainé 171 et de benjamin 186.

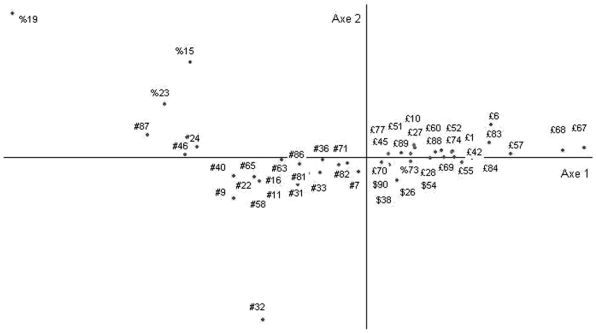


FIG 12. — Représentation des classes 171-1(%),171-2($),186-1(#),186-2(£) issues du nœud 188.

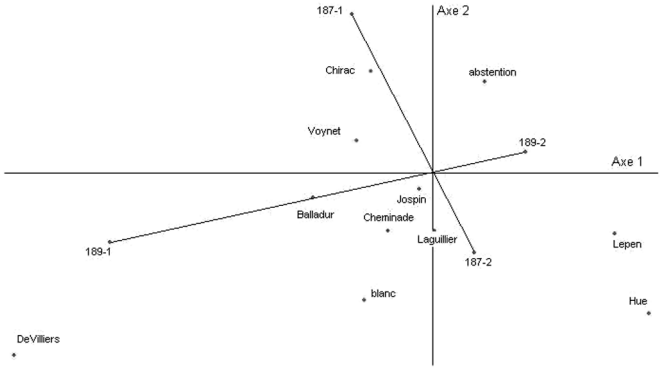


FIG 13. — Plan factoriel associé au nœud 190 d'ainé 189 et de benjamin 187.

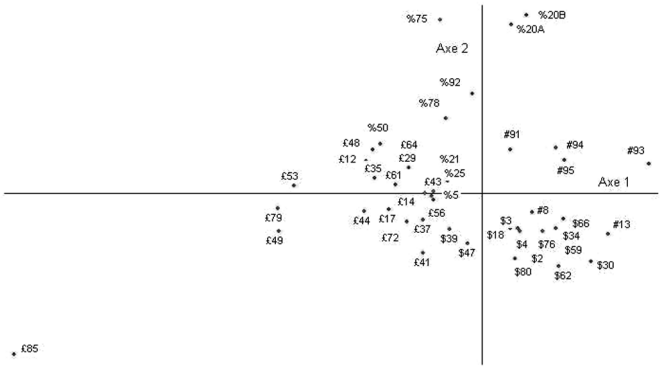


FIG 14. — Représentation des classes 189-1(£),189-2(#),187-1(%),187-2($) issues du nœud 190.

Références

[1] BENZECRI J.P. (1976). L'Analyse des Données (Volumes I and II). Dunod, Paris.

[2] BRUYNOOGHE M. (1978). Large data set clustering methods using the concept of space contraction. Compstat. 3, Physika Verlag, Vienna, pp 239-245.

[3] DAZY F., LE BARZIC J.F. (1996). L'analyse des données évolutives. Technip.

[4] DENIMAL J.J. (2000). Correspondances hiérarchiques : une nouvelle approche. XXXII^{ieme} Journées de Statistiques, 15-19 mai 2000. Fès, Maroc.

[5] DENIMAL J.J. (2001). Hierarchical factorial analysis. 10th International Symposium on Applied Stochastic Models and Data Analysis. 12-15 juin 2001. Compiègne.

[6] DENIMAL J.J., CAMIZ S. (2001). Exact conditional tests for a reciprocal interpretation of hierarchical classifications built on a txwo-way contingency table. Metron,Vol. LIX, n°. 3-4, pp 157,178.

[7] DENIMAL J.J. (2007). Classification factorielle optimisée d'un tableau de mesures. Revue de Statistique Appliquée (à paraître).

[8] DIDAY E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes. Revue de Statistique Appliquée, Vol.19, n°2, pp 19,34.

[9] GAIL M., MANTEL N. (1977). Counting the number of r×c contingency tables with fixed margins. Journal of the American Statistical Association, Vol. 72, n°360, pp 859,863.

[10] GOVAERT G.,(1984). Classification simultanée de tableaux binaires. Data Analysis and Informatics, 4, Diday et al. Eds, North Holland,pp 223,236.

[11] JUAN J. (1982). Classification automatique hiérarchique selon les voisins réciproques. Les cahiers de l'analyse des données, Vol 7, n°2.

[12] LANCASTER H.O. (1969). The Chi-squared distribution. John Wiley and Sons, New York.

[13] LEBART L., MORINEAU A., PIRON M. (1995). Statistique exploratoire multidimensionnelle.Dunod, Paris.

[14] Mehta C.R., Patel N.R. (1983). A network algorithm for performing Fisher's exact test in r×c contingency tables. Journal of the American Statistical Association, Vol. 78, n°382, pp 427,434.

[15] Patefield W.M. (1981). An efficient method of generaing random r×c tables with given row and column totals. Applied Statistics, Vol. 30, pp 91,97.

[16] Plackett R.L. (1981). The analysis of categorical data. Second Edition, Griffin, London.

[17] Ward J.H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, pp 236-244.